University of Alberta

Use of model organisms and phylogenetic analysis to characterize the role of *CECR1* in cat eye syndrome

by

Stephanie Ann Maier ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in

Molecular Biology and Genetics

Department of Biological Sciences

Edmonton, Alberta

Fall 2005

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

0-494-08688-2

Your file Votre référence
ISBN:
Our file Notre reference
ISBN:

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Dedication

This thesis is dedicated to my husband Mike, who has never known what it is like for his wife not to be in school. This is the end of one very long chapter in my life, and the beginning of the rest of our lives together!

Abstract

Cat eye syndrome (CES) is a rare genetic disorder associated with the partial duplication of chromosome 22q11, involving defects of the eyes, ears, heart, kidney, and urogenital tract. The *CECR1* gene is a very promising candidate for the production of CES features when overexpressed, based on both its expression profile and sequence similarity to growth factors. RNA *in situ* hybridization showed that *CECR1* is faintly expressed throughout most of the embryo, and specifically in the developing heart and kidney, two tissues affected in CES. These experiments also uncovered an antisense transcript to *CECR1*, which was confirmed using Northern blot analysis and RT-PCR. The presence of this antisense transcript suggests that *CECR1* is regulated at the post-transcriptional level. A second transcript was also identified within the *CECR1* genomic region, called *CECR1* variant 2 (*CECR1v2*), making the study of *CECR1* more complex.

The CECR1 protein is part of a group of proteins with similarity to ADA, called the Adenosine Deaminase-related Growth Factors (ADGF, known as CECR1 in vertebrates). Studies in Drosophila revealed six ADGF homologues, with differential expression patterns. Protein sequences related to the ADGF and ADA genes were predicted from as many taxa as possible. The phylogenetic relationship of these gene products was determined using parsimony and Bayesian methods, and a novel paralogue was discovered, termed ADA-like (ADAL). An analysis of conserved residues showed that both the ADGF and ADAL subgroups have all the required residues for ADA activity. The availability of genomic data for the members of this family allowed the reconstruction of intron evolution within the phylogeny. Overall, ADA activity is clearly more complex than once thought, perhaps involving a delicately balanced pattern of temporal and spatial expression of a number of paralogous proteins.

Acknowledgments

I would like to thank my supervisor, Heather McDermid, for accepting me into her lab, teaching me so much, and supporting me throughout this entire process. Thanks also to my committee, Frank Nargang and Moira Glerum, for helpful suggestions.

Thanks to Song Hu, Wayne Materi, and Bryan Crawford for teaching me so much about techniques and science in general. My undergraduate project students, Hannah Cheung, Jon Staav, and Julia Galellis, each taught me a bit about myself while I tried to guide them in their projects. Thanks also to Twila Yobb, Rezika Zurch, Fang Yang, Nic Fairbridge, and summer students Katie Kessler and Cheryl Johnson, who all helped me with some aspects of this work. Polly Brinkman-Mills, Heather Wilson, and Isabelle Mousseau were great friends both in the out of the lab. All the other McDermid lab members were great to work with, and made the time here enjoyable.

Thanks to Randy Mandryk and especially Jack Scott for helping me with tissue embedding and microscopy work, and for tips on formatting my figures. Sara Zalik, Tania Attie-Bitach, and Louis Honore helped me to identify embryonic structures in the pig and human embryo slides. Thanks also to John Locke and Lynn Podemski for helping me understand Drosophila genetics. Special thanks to Warren Gallin for guidance and helpful discussions throughout both the phylogenetic and Western analyses.

The completion of this thesis would not have been possible without my husband Mike, who was always my number one supporter. His understanding, patience and love were endless throughout this undertaking. My family back home was also always supportive of me, despite not really knowing what I was doing!

The Canadian Institutes of Health Research, the Alberta Heritage Foundation for Medical Research, and the University of Alberta provided financial assistance in the form of graduate studentships and tuition assistance throughout my project.

Table of Contents

Chapter 1: Introduction **1**

Congenital defects associated with chromosome 22q11.2 .... **1**

Cat eye syndrome **2**
  Features of cat eye syndrome patients 2
  Molecular pathology 3
  Delineation and sequencing of the CES critical region 4
  Candidate genes within the CES critical region 4

***CECR1* as a candidate gene for CES** **6**
  CECR1 discovery and transcript properties 6
  Expression pattern of *CECR1* 7
  Sequence homology to growth factors 7

Growth factors **8**

Adenosine deaminase (ADA) **9**
  Three isoforms of ADA 9
  Protein structure and active site residues of ADA1 13

ADGF family members **14**

Phylogenetic inference **16**

Intron positions and the early/late debate **19**

Research objectives **22**

Chapter 2: Materials and Methods **29**

Isolation of nucleic acids **29**
  Plasmid DNA 29
  Genomic DNA 29
  DNA embedded in agarose 29
  RNA isolation 30

Preparation of BAC clone to create *CECR1* transgenic mice **30**

DNA probe preparation **32**

Southern analysis **32**

Northern analysis **33**

RT-PCR **34**

PCR **35**

Sequencing **35**

Screening a zebrafish cDNA library **36**

Western analysis **37**

Antibody production to CECR1 ... 37
Protein sample preparation and quantification ... 38
Western gel electrophoresis and transfer ... 39
Western detection of proteins ... 39
Competition assay ... 40

ELISA ... 40

Protein A pull-down experiment ... 41
Preparation of Protein A beads ... 41
Large-scale protein A precipitation ... 41

Cell culture ... 42

***In situ* hybridization ... 42**
Embryo collection and storage ... 42
Preparation and testing of digoxigenin-labeled RNA probes ... 43
Whole mount *in situ* hybridization ... 44
Slide *in situ* hybridization ... 46

Sequence analysis and computer software ... 47
Gene discovery, prediction and annotation tools ... 47
Multiple alignment and phylogenetic analysis ... 48
Mapping of intron positions ... 49

Chapter 3: Results ... 55

Defining the proximal CES critical region: Genomic annotation of *IL-17R* and *CECR1* ... 55
Analysis of the 3' end of *IL-17R* ... 55
Discovery and characterization of *CECR1* variant 2 (*CECR1v2*) ... 56

Overexpression of human *CECR1* in a transgenic mouse model ... 59
Characterization of the sequences present on human BAC 609c6 ... 59
Production of transgenic founders ... 61
Expression of human *CECR1* in the transgenic mice ... 61
Phenotypic observations and mutation analysis of the transgene ... 62

Production and analysis of human CECR1 antibodies ... 63
Rabbit anti- recombinant human CECR1 antibody production (0A1) ... 63
Rabbit anti- human CECR1 peptide antibody production (2F6, 2F2, and 2F1, 2F3) ... 64
Protein A pull-down of CECR1 from cell lysates ... 65

Expression analysis of *CECR1* ... 66
Zebrafish whole mount *in situ* hybridization ... 66
Pig *in situ* hybridization ... 67
Human *in situ* hybridization ... 69
Confirmation of the antisense transcript ... 72

Gene structure and expression patterns of the Drosophila *ADGF* genes ... 74
Identification and sequencing of genes ... 74
Genomic structure of the Drosophila homologues ... 75

Conservation of intron positions 77
Expression analysis of the six Drosophila *ADGF* genes 77

Phylogenetic analysis of the ADGF subfamily 79
Identification and sequencing of preliminary protein sequences 79
No mouse CECR1 homologue exists 80
Discovery of the ADAL paralogues 81
Identification of protein sequences in silico for use in the phylogenetic analysis 81
Prediction of signal peptides 84
Alignment of protein sequences 85
Initial phylogenetic inference 87
Focused analysis of the ingroup and aspects of MrBayes analyses 88
Observations from the Bayesian analysis 90
Parsimony analysis 93
Intron evolution in the ingroup 94

Chapter 4: Discussion 141

Human *IL-17* Receptor 141

Interpretation of animal studies in *CECR1* overexpression 141

CECR1 antibodies only reliably detect the control samples 143

Striking similarities between CECR1 and ADA2 144

RNA *in situ* hybridization experiments are consistent with CES features 146

Discovery of an antisense transcript indicates that *CECR1* may be regulated post-transcriptionally 149
Proof of existence for the antisense transcript 149
The *CECR1* sense and antisense transcripts are differentially expressed 150
Type of overlap involved with the antisense transcript 151
Possible mechanisms involved in regulation of the *CECR1* transcript 152

There are six *ADGF* homologues in Drosophila 154
Gene structure and expression pattern individualize the *ADGF* genes 154
Gene orientations and theories of regulation 155
Theories of gene duplication and divergence 156
Possible function of the six Drosophila homologues 157

Phylogenetic analysis reveals an evolutionary relationship between the ADGF, ADAL, and ADA subfamilies 159
Parameters within the Bayesian analysis and differences from Maximum parsimony 159
Conservation of ADA active site residues 160
Patterns revealed in the phylogenetic analyses 162
Missing members in different organisms 163
The presence of multiple ADGF members in some organisms suggests exploitation of alternate functions 164

Proof for the introns-late aspect of the new synthetic theory of introns 166

Since *CECR1* is missing in mouse, is *ADAL* or *ADA* providing compensation? 167
Human *CECR1* as a candidate for cat eye syndrome 168
The involvement of *CECR1* variant 2 in CES 170
Conclusions 173
Future Directions 173
  Use of model organisms for deletion/duplication studies 173
  Biochemical properties of CECR1 175
  Protein localization and binding assays 175
  Increased confidence in the phylogenetic analysis with more protein sequences 176
  Further characterization of *CECR1v2* 176
  Characterization of the *CECR1* antisense transcript 178
Significance of this work 179

References 181

Appendix 197

List of Tables

Table 2-1. Primer sequences used in the study 51

Table 3-1. Names, accession numbers, and predicted signal peptides of proteins used in the study. 130

Table 3-2. Summary of temperature settings and acceptance values for individual MrBayes analyses of the ingroup alignment. 135

Table 4-1. Predicted molecular weight (mw) of human proteins in the adenyl-deaminase family. 180

List of Figures

Figure 1-1. Complement of genetic material from chromosome 22 in a typical CES patient. 23

Figure 1-2. Putative genes identified in the CES critical region and syntenic region in mouse. 24

Figure 1-3. Northern analysis of human *CECR1*. 25

Figure 1-4. Schematic depiction of the proteins involved in ADA activity. 26

Figure 1-5. Schematic drawing of the reaction mechanism of adenosine deaminase..... 27

Figure 1-6. Reaction mechanisms involved in adenine nucleotide catabolism. 28

Figure 2-1. Amino acid substitution matrix used for maximum parsimony analysis..... 54

Figure 3-1. Genetic structure of *IL-17R*. 98

Figure 3-2. Expression analysis of human *IL-17R*. 99

Figure 3-3. Genomic structure and predicted protein sequences of human *CECR1*..... 100

Figure 3-4. Expression analysis of the different *CECR*1 isoforms. 101

Figure 3-5. Characterization of human BAC 609c6. 102

Figure 3-6. Expression analysis of human *CECR1* in transgenic mice. 103

Figure 3-7. Hypothetical CECR1 protein sequences and locations of antigens. 104

Figure 3-8. Western analysis of spleen extracts using the 0A1 antibody. 105

Figure 3-9. Competition assay using the HIDExp1 recombinant protein. 106

Figure 3-10. Expression analysis of zebrafish *CECR1-1*. 107

Figure 3-11. Locations of *in situ* hybridization probes in the pig and human *CECR1* genes. 108

Figure 3-12. Whole-mount *in situ* hybridization of day 20 pig embryos using various probes. 109

Figure 3-13. *In situ* hybridization of day 20 pig embryo sections using various probes. 110

Figure 3-14. *In situ* hybridization of day 28 pig embryo sections using various probes. 111

Figure 3-15. *In situ* hybridization of day 31 pig embryo sections using various probes. 112

Figure 3-16. RNA *in situ* hybridization of human fetal day 34 sections using the CECR1-1AS or CECR1-1S probe. 113

Figure 3-17. RNA *in situ* hybridization of human fetal day 34 sections with the CECR1-1AS or CECR1-4AS probe. 114

Figure 3-18. RNA *in situ* hybridization of human fetal day 34 sections using the CECR1-4S probe. 115

Figure 3-19. RNA *in situ* hybridization of human fetal day 47 sections using various CECR1 probes. 116

Figure 3-20. RNA *in situ* hybridization of human fetal week 8.5 kidney sections. 117

Figure 3-21. RNA *in situ* hybridization of human fetal week 8.5 lung sections. 118

Figure 3-22. RNA *in situ* hybridization of human fetal week 8.5 sections of the stomach. 119

Figure 3-23. RNA *in situ* hybridization of human fetal week 10.7 liver sections. 120

Figure 3-24. RNA *in situ* hybridization of human fetal week 10.7 metanephric kidney sections. 121

Figure 3-25. Expression profile of pig *CECR1* transcripts. 122

Figure 3-26. Expression profile of the human *CECR1* antisense transcript. 123

Figure 3-27. Summary of results for the *CECR1* antisense transcript. 124

Figure 3-28. Drosophila *ADGF* homologues. 125

Figure 3-29. Summary of RT-PCR results in the 75A region. 126

Figure 3-30. Schematic depiction of the intron alignment between the Drosophila *ADGF*s and human *CECR1*. 127

Figure 3-31. Expression analysis of the Drosophila *ADGF* genes during development. 128

Figure 3-32. Alignment and conserved domains in the adenyl-deaminase family. 133

Figure 3-33. Initial analysis of the five protein subfamilies using MrBayes. 134

Figure 3-34. Convergence to stationarity during Bayesian phylogenetic analysis of the ingroup for individual MrBayes runs. 136

Figure 3-35. Phylogenetic analysis of the ingroup using MrBayes. 137

Figure 3-36. Phylogenetic analysis of the ingroup using Maximum parsimony. 138

Figure 3-37. Evolution of introns within the ingroup. 139

Figure 3-38. Maximum parsimony analysis of intron positions. 140

Figure A1. Vector map and multiple cloning site of the pGEM-T Easy vector (Promega) used for cloning PCR products. 197

Figure A2. Vector map and multiple cloning site of pBluescript II SK- (Stratagene). 198

Figure A3. Vector map and multiple cloning site for the pRSET A (Invitrogen) expression vector. 199

Figure A4. Vector map of pZErO-2 (Invitrogen). 200

List of Abbreviations

aa	amino acids
ADA	adenosine deaminase
ADAL	adenosine deaminase - like
ADE	adenine deaminase
ADGF	adenosine deaminase-related growth factor
AMPD	adenosine monophosphate deaminase
ATP	adenosine triphosphate
BAC	bacterial artificial chromosome
BLAST	basic local alignment search tool
bp	base pairs
cDNA	complementary DNA
CECR1	cat eye syndrome critical region gene 1
CES	cat eye syndrome
DCF	2'-deoxycoformycin
DGS	DiGeorge syndrome
DIG	digoxigenin
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
dATP	deoxyadenosine triphosphate
dNTP	deoxynucleotide triphosphate
dpf	days post fertilization
dsDNA	double-stranded DNA
EHNA	(+)-erythro-9(2-S-hydroxy-3-R-nonyl)adenine
EST(s)	expressed sequence tag(s)
GFP	green fluorescent protein
GST	glutathione-S-transferase
hpf	hours post fertilization
IDGF	insect-derived growth factor
IL-17	interleukin-17

IL-17R	interleukin-17 receptor
kb	kilobases (RNA) or kilobase pairs (DNA)
kDa	kilodaltons
LCR22	low-copy repeat on chromosome 22
LINE	long interspersed nuclear element
Mb	magabases (RNA) or megabase pairs (DNA)
MDGF	mollusk-derived growth factor
MHC2b	myosin heavy chain 2b
MP	maximum parsimony
mRNA	messenger RNA
MSI	male-specific IDGF
nr	database of non-redundant sequences
OMIM	online mendelian inheritance in man
ORF	open reading frame
PAC	P1 bacteriophage-based artificial chromosome
PCR	polymerase chain reaction
PFGE	pulsed-field gel electrophoresis
pfu	plaque forming units
poly(A)+	polyadenylated RNA
RACE	rapid amplification of cDNA ends
RNA	ribonucleic acid
RNase	ribonuclease
RT-PCR	reverse transcriptase - polymerase chain reaction
SCID	severe combined immunodeficiency syndrome
ssDNA	single-stranded DNA
TAPVR	total anomalous pulmonary venous return
T_M	melting temperature
TOF	Tetralogy of Fallot
TSGF	tsetse salivary growth factor
UTR	untranslated region
VCFS	velocardiofacial syndrome

Chapter 1: Introduction

Loss of pregnancy due to aneuploidy is a common occurrence. The trisomy of only three chromosomes (13, 18, and 21) has been routinely observed, and only those with trisomy 21 survive past the neonatal period (reviewed in Oyler et al., 2004). In order to understand the molecular pathology of aneuploidy events it is advantageous to evaluate candidate genes within a small region of duplication. This thesis describes the characterization of *CECR1*, a candidate gene within the duplicated region of 22q11.2 associated with cat eye syndrome.

Congenital defects associated with chromosome 22q11.2

Chromosome 22 is the second smallest human autosome, comprising 1.6 – 1.8% of total genomic DNA (Morton, 1991). It is an acrocentric chromosome, and the short p-arm encodes tandem repeats including α-satellites and ribosomal RNA genes. The long q-arm of chromosome 22 is gene rich compared to other chromosomes (reviewed in Dunham et al., 1999), and as such there are numerous genetic diseases associated with the long arm of chromosome 22. The 22q11.2 region in particular hosts a number of congenital chromosomal rearrangements, due to the presence of several unstable low-copy repeats called LCR22s (Edelmann et al., 1999). Both inter- and intra-chromosomal recombination events can occur to produce deletions and/or duplications (Edelmann et al., 1999). Congenital rearrangements associated with the 22q11.2 region include cat eye syndrome (CES), 22q11.2 deletion syndrome, and der(22) syndrome. Cat eye syndrome is the subject of this thesis and will therefore be discussed in detail in the next section.

The 22q11.2 deletion syndrome (also known as DiGeorge syndrome (DGS) [OMIM 188400], or velocardiofacial syndrome (VCFS) [OMIM 192430]) is associated with the hemizygous partial deletion of chromosome 22q11.2, and occurs once in every 4000 live births (reviewed in Baldini, 2003). The size of the deletion is most often 3 Mb, governed by two LCR22s, although a nested 1.5 Mb deletion is also possible due to the presence of an internal LCR22 (Baldini, 2003). Common features include cardiac outflow tract abnormalities, absence or hypoplasia of the thymus and parathyroid glands, T-cell deficits, cleft palate, facial anomalies, hypocalcemia, and mental retardation, although the phenotype is highly variable (reviewed in Yagi et al., 2003). The gene thought to be responsible for the majority of defects associated with this syndrome, *TBX1*, encodes a transcription factor of the T-box family (Yagi et al., 2003). The concurrent deletion of any number of the approximately 24-30 other genes in this region, however, is required to produce the full phenotype. The microduplication of the exact same region has also been discovered, which results in a milder phenotype that can often go undetected, and therefore may be just as common as the deletion (Ensenauer et al., 2003). In fact, this region is often duplicated in CES patients, without any obvious additional phenotype in the small sample size examined (McTaggart et al., 1998; McDermid and Morrow, 2002).

Der(22) syndrome is a rare disorder associated with multiple congenital abnormalities (Shaikh et al., 1999). Patients carry a supernumerary t(11;22)(q23.3;q11.2) chromosome, as a result of a mis-segregation event in a balanced carrier, and are therefore trisomic for 22pter-q11.2 and 11q23.3-qter (Zackai and Emanuel, 1980; Fraccaro et al., 1980). The presence of this extra chromosomal material results in a distinct phenotype, consisting of severe mental retardation, preauricular tags, ear anomalies, cleft or high-arched palate, micrognathia, microcephaly, kidney abnormalities, heart defects, and genital abnormalities in males (reviewed in Shaikh et al., 1999). There is some phenotypic overlap between der(22) syndrome and CES, since the region that is trisomic in der(22) syndrome overlaps with the interval that is triplicated in most cat eye syndrome patients (Funke et al., 1999).

Cat eye syndrome

Features of cat eye syndrome patients

Cat eye syndrome (CES, OMIM 115470) is a rare (incidence of between 1:50,000 and 1:150,000) human genetic disorder associated with the duplication of a region of chromosome 22q11 (McDermid et al., 1986). The major clinical features of CES include preauricular skin tags and/or pits, anal atresia (with or without fistula), kidney/urogenital malformations, ocular coloboma (of the iris and/or retina), and cardiac defects (Schinzel et al., 1981; Rosias et al., 2001). Minor features can include downslanting palpebral fissures, hypertelorism, orthopedic deformities, low set ears, abdominal malformations, and mild to moderate mental retardation (Schinzel et al., 1981; Rosias et al., 2001). The phenotype is highly variable in that no feature is present in all individuals, and the severity varies enough that some mildly affected patients probably remain undetected. In fact only 9 patients out of 105 cases reviewed from the literature showed all the major clinical features (Rosias et al., 2001).

Molecular pathology

The clinical diagnosis of CES is confirmed by the cytological finding of a marker chromosome, usually in the form of an isodicentric bisatellited chromosome derived from 22pter-q11.2 (McDermid et al., 1986). The presence of the marker chromosome results in the partial tetrasomy of the entire p arm and the most proximal part of the q arm (Figure 1-1). The marker can be symmetric or asymmetric, depending on the breakpoint used by each of the inverted chromosomal pieces (Mears et al., 1994). The same two LCR22s that are responsible for generating the 3 Mb deletion in the 22q11.2 deletion syndrome also cause the rearrangements to form the two types of CES marker chromosomes (McTaggart et al., 1998). If the CES chromosome is symmetrical, with both pieces originating from the proximal breakpoint, it is referred to as a type I CES chromosome. Type II CES chromosomes may be asymmetrical or symmetrical, where one or both breakpoints occur in the distal LCR22 (McTaggart et al., 1998). Recently, a type III CES chromosome was described that was bisatellited (a chromosome 22 p-arm located on each end of the marker) but only contained one copy each of the chromosome 22 centromere and CES critical region, and had its breakpoint at 22q12.3 (Bartsch et al., 2005). Since this patient was mosaic and the only one found with this novel marker, however, it is not known how common or significant this finding is.

Some patients, rather than having a marker chromosome, harbour an interstitial duplication that results in trisomy of the critical region. Patients “LW” (Reiss et al., 1985) and “SK” (Knoll et al., 1995) fall within this category, as well as a third patient (Lindsay et al., 1995), and each shows a partial CES phenotype not unlike most CES patients with four copies of the region. Recently, a fourth CES interstitial duplication patient was discovered that exhibited all of the major and some minor CES symptoms (Meins et al., 2003). The CES marker chromosome can also take the form of a ring, in which one or two extra copies of the 22pter-q11.2 region are present. A dicentric ring chromosome containing two extra copies of the CES region was thought to be responsible for the phenotype observed in a patient with many of the features of CES (Mears et al., 1995). These examples illustrate that the severity of the phenotype can not be directly correlated to the extent of duplicated material on chromosome 22, nor to the number of copies (reviewed in Rosias et al., 2001).

Delineation and sequencing of the CES critical region

Two patients that exhibited almost all of the cardinal features of CES were used to define the CES critical region, the duplication of which is required to produce the CES phenotype. One patient (CM15) who showed all the major features of the syndrome had an unusually small supernumerary dicentric ring chromosome consisting of the first 2 Mb of 22q (Mears et al., 1995). An interstitial duplication in patient SK (Knoll et al., 1995) narrowed the critical region further to approximately 1 Mb of the distal half of the original 2 Mb (H. McDermid, unpublished). SK exhibited all of the features except anal atresia and ocular coloboma, which suggested that the genes responsible for these phenotypes might be located in the proximal 1 Mb region, or more likely could be explained by the phenotypic variability of the syndrome (Knoll et al., 1995).

A 1.5 Mb region containing the CES critical region was cloned into a set of bacterial and P1-based artificial chromosomes (BACs/PACs) (Johnson et al., 1999). A minimal tiling path of clones was chosen to be sequenced by Bruce Roe at the University of Oklahoma, and the data was published along with the rest of the chromosome 22 sequence (Dunham et al., 1999).

Candidate genes within the CES critical region

Using various techniques including exon trapping, sequence annotation, EST analysis, comparative genomics, and RT-PCR, 14 putative human genes (Figure 1-2) were discovered in and around the CES critical region (Footz et al., 2001). Concurrently, the syntenic region on mouse chromosome 6 was sequenced, and 10 putative orthologues to the 14 human genes were uncovered (Footz et al., 2001). The human genes were each evaluated for their potential as a CES candidate gene. Candidate genes are those predicted to be dosage sensitive and, upon formation of the bisatellited marker chromosome, become overexpressed to give rise to at least some of the features of CES (Footz et al., 2001). Dosage sensitive genes might be those that encode proteins such as transcription regulators, growth factors, receptors, structural proteins (Fisher and Scambler, 1994), chromatin proteins, and members of signal-transduction cascades (Birchler et al., 2005). Gene expression in the tissues affected in CES patients was also considered when assigning candidacy (Footz et al., 2001).

Two of the fourteen putative genes were already identified before the CES critical region was mapped and sequenced. *ATP6E* was the first gene mapped to the region, and codes for the epsilon subunit of vacuolar ATPase (Baud et al., 1994). Due to its ubiquitous pattern of expression and the fact that similar genes are involved in autosomal recessive diseases, *ATP6E* was not considered to be a good candidate for CES and was not studied further (Footz et al., 2001). The *IL-17R* gene was also previously mapped to the 22q11 region, and encodes the receptor of the IL-17 cytokine (Yao et al., 1997). *IL-17R* is expressed globally (Yao et al., 1997), while its ligand is only expressed in T-cells (Yao et al., 1995), suggesting that the ligand is the limiting factor and that changes in dosage of the *IL-17R* gene might not cause any effect. When combined with the fact that CES patients do not exhibit any overt immune system abnormalities (Rosias et al., 2001), the *IL-17R* gene was given a low priority as a candidate for production of CES features (Footz et al., 2001).

Of the remaining twelve genes discovered in the CES critical region, three were considered very good candidates for production of CES features, and were therefore chosen for further study: *CECR1*, *CECR2*, and *CECR6*. *CECR1* is the focus of this thesis and will be described in detail below, and in the body of the thesis. *CECR2* encodes a member of a chromatin remodeling complex, and could therefore be dosage sensitive (Banting et al., 2005). Since it is expressed in neural tissue as well as in the developing eye, *CECR2* might play a role in the mental retardation and ocular coloboma observed in CES patients (Banting et al., 2005). The *CECR6* gene might also be a good candidate, since it is expressed in a number of tissues consistent with CES features, including fetal brain and kidney, and adult heart (Footz et al., 2001). It encodes an interesting protein that contains multiple amino acids runs and multiple transmembrane domains (Mousseau, 2005), suggesting that it might function as a receptor that may be dosage sensitive.

The remaining genes were considered less promising at the time and were therefore not studied further due to limited lab resources (Footz et al., 2001). Two genes, *CECR7* and *CECR8*, were localized to the pericentromeric region and appeared to be aberrant non-functional transcripts (Footz et al., 2001; Bridgland et al., 2003). *CECR3*, *CECR4*, and *CECR9* had either incomplete gene structures, or were not predicted to encode a functional protein. *CECR5* and *SLC25A18* were not expected to be dosage sensitive, since *CECR5* showed similarity to an enzyme, and genes with similarity to *SLC25A18* cause autosomal recessive disorders when mutated (Footz et al., 2001). Finally, *MIL1* probably does and *BID* actually does lie outside the CES critical region breakpoint as defined by patient CM15, which has not been mapped completely, but either gene could confer some subtle effects in patients that have this region duplicated (Footz et al., 2001). Overall, it is not clear whether CES is caused by the duplication of one or many genes within the CES critical region. As such, although the preceding genes were considered low priority based on the information available at the time, each gene must eventually be characterized fully to determine its role in CES.

***CECR1* as a candidate gene for CES**

CECR1 *discovery and transcript properties*

CECR1 (Cat Eye syndrome Critical Region gene 1) is a particularly promising CES candidate that was originally isolated using exon trapping in the McDermid lab (Riazi et al., 2000). Sequencing of IMAGE clone 54445 (accession # AA348024) combined with 5' RACE revealed that the *CECR1* transcript is 3941 bp in length and its nine exons and eight introns span approximately 30.5 kb of chromosomal DNA. The last 2.2 kb of *CECR1* are composed almost entirely of *Alu* and LINE repeat sequences contained within the 3' UTR. The remainder of the sequence comprises an open reading frame of 1536 bp, which encodes 511 amino acids (Riazi et al., 2000). Although an in-frame upstream stop signal has not yet been located, it seems probable that the entire coding region is present in this open reading frame, due to the presence of a putative signal peptide involving the first 29 amino acids. Also, the 3.9 kb cDNA transcript approaches the 4.4 kb transcript size (see below) if the as yet unknown 5' UTR is taken into account.

Expression pattern of CECR1

Northern analysis has determined that *CECR1* is expressed in various adult and fetal tissues important for CES features (Riazi et al., 2000). Two differentially expressed *CECR1* transcripts were discovered, approximately sized 4.4 kb and 3.5 kb (see Figure 1-3). The larger band is expressed in human placenta, adult heart, lung and lymphoblast, and fetal lung and liver. The smaller 3.5 kb band is expressed in adult heart, kidney, pancreas, and lymphoblast, and fetal lung and kidney. The 4.4 kb band corresponds to the full length *CECR1* transcript, and although not confirmed, it was thought that the smaller 3.5 kb band might result from alternative polyadenylation (Riazi et al., 2000).

Embryonic expression was studied further using *in situ* hybridization of a *CECR1* RNA probe on day 35 human embryo sections (Riazi et al., 2000). Expression of the *CECR1* transcript was found in the outflow tract and atrium of the heart, the VII/VIII cranial nerve ganglion (precursor to the facial and acoustic ganglions, respectively), the developing notochord, and the placenta. This pattern of expression fits well with the tissues affected in CES patients.

Sequence homology to growth factors

Blast searches using the CECR1 protein sequence revealed significant sequence similarity to a number of putative growth factors from other organisms (Riazi et al., 2000). CECR1 showed 38% amino acid identity to insect-derived growth factor (IDGF) from *Sarcophaga peregrina* (Homma et al., 1996), 39% identity to mollusk-derived growth factor (MDGF) from *Aplysia californica* (Sossin et al., 1989), and 33% and 40% identity to *Glossina morsitans* salivary gland growth factors TSGF-1 and -2 (Li and Aksoy, 2000), respectively. All of these proteins also share sequence similarity to adenosine deaminase (ADA) in their C-terminal portions, with conservation of the ADA active site residues, suggesting that the active domain may be important in the function of these proteins. These growth factors will be described in more detail below, in the section concerning the ADGF family members.

Overall, the localization of *CECR1* in the CES critical region, its expression in tissues affected in CES patients, and its putative role in growth regulation make it an attractive candidate gene for a role in at least some CES features.

Growth factors

Growth factors are usually thought of as ligands that function by interacting with a specific receptor on the surface of the cell, which causes conformational changes in the receptor and results in a signal transduction cascade that causes changes in gene expression. For example, TGF-beta signaling is initiated by the ligand binding to a membrane-associated receptor complex that has serine/threonine kinase activity. This receptor complex phosphorylates specific proteins that then transduce the ligand-activated signal to the nucleus (reviewed in Cheng and Grande, 2002). The epidermal growth factor (EGF)-related peptides bind various ErbB receptors, inducing dimerization and phosphorylation of specific tyrosines in the receptors cytoplasmic region. These phosphorylated residues serve as docking sites for a variety of signaling molecules whose recruitment stimulates intracellular signaling cascades, which ultimately control cell growth (reviewed in Holbro and Hynes, 2004). However, other methods of stimulating growth exist, including the regulation of low molecular weight substances such as adenosine, which is an important signaling molecule.

Adenosine binds to one of four adenosine receptor subtypes (A_1, A_{2a}, A_{2b}, and A_3) on the cell surface (Franco et al., 1998). These seven transmembrane receptors belong to the superfamily of G protein-coupled receptors involved in cell signaling (Franco et al., 1998). Extracellular adenosine is a modulator that acts through these adenosine receptors to produce different physiological effects involving cell proliferation and migration, angiogenesis, neurotransmission, lymphocyte function, blood pressure, heart rate, and renal function, (reviewed in Akalal et al., 2004). The effect produced by adenosine may change according to which receptors are present on the cell surface. For example, binding of adenosine to A_1R inhibits the activity of membrane adenylyl cyclase, whereas binding to one of the A_2 receptors stimulates the activity of adenylyl cyclase (reviewed in Jacobson et al., 1999). Adenosine has been shown to inhibit the growth of rat vascular smooth muscle cells (Dubey et al., 1996) and rat cardiac fibroblasts (Dubey et al., 1997). Human endothelial cells were stimulated to proliferate in the presence of adenosine, but were inhibited by high adenosine concentrations, while these effects were not observed when human lung fibroblasts were used (Ethier et al., 1993). This suggests that adenosine effects are cell-type specific and are dependent on adenosine concentrations. Overall, these examples stress that the concentration of extracellular adenosine must be tightly regulated.

Adenosine deaminase (ADA)

Three isoforms of ADA

Adenosine deaminase (ADA, OMIM 102700) catalyzes the deamination of adenosine and 2-deoxyadenosine to inosine and 2-deoxyinosine, respectively. Human ADA activity is highest in the thymus but has been observed in all human tissues, due to at least three isoforms: ADA1, ADA1+CP, and ADA2 (reviewed in Hirschhorn and Ratech, 1980; Figure 1-4). The two ADA1 protein forms can be specifically inhibited by (+)-erythro-9(2-S-hydroxy-3-R-nonyl)adenine (EHNA), while all three forms are inhibited by 2'-deoxycoformycin (DCF) (Niedzwicki and Abernethy, 1991).

The *ADA1* gene is located on chromosome 20q12-13, and encodes a 363 amino acid protein of approximate molecular weight 41 kDa. ADA1 deficiency results in one type of Severe Combined Immune Deficiency (ADA-SCID) in which patients show reduced or absent B- and T-cells and therefore have a nonfunctioning immune system (reviewed in Hershfield, 2003). It is thought that the severe lymphopenia is largely due to the conversion of ADA substrates to dATP, the accumulation of which inhibits a key enzyme in DNA synthesis, ribonucleotide reductase, and stabilizes pro-apoptotic complexes (Hershfield, 2003). ADA1 activity has been found to be ubiquitous in a number of vertebrate studies, and although its specific activity is very low in erythrocytes, it is the only species of ADA present in erythrocytes (Hirschhorn and Ratech, 1980). ADA1 is also more prevalent in tissues such as spleen and stomach that exhibit high specific ADA activity (Van der Weyden and Kelley, 1976; Meng et al., 1997). Although Northern analysis on human patient lymphoblasts showed the expected 1.6-1.8kb transcript size (Daddona et al., 1985), no other studies of the RNA distribution in human tissues have been published.

ADA1+CP is a 280 kDa protein complex composed of two ADA1 enzymes bound together by a combining protein (CP) that has been identified as CD26 (also known as dipeptidyl peptidase IV (DPPIV)) (Franco et al., 1998). In human tissues with low specific ADA activity such as lung and kidney, this large form of ADA1 predominates, but it has also been found in liver and intestine (Van der Weyden and Kelley, 1976). CD26 is a membrane glycoprotein that cleaves dipeptides from the N-terminus of polypeptides with proline in the penultimate position (Fleischer, 1994). Within this complex, the ADA1 enzyme has been termed "ecto-ADA" since it is located on the outside of the cell (Figure 1-4). The binding of ecto-ADA to CD26 does not interfere with the enzymatic activity, as it has been shown that ecto-ADA is effective in degrading extracellular adenosine (Franco et al., 1997). Ecto-ADA has been found in nearly all cell types, but it is not necessarily present in each cell of that tissue (reviewed in Franco et al., 1998). CD26 is found on a variety of different cell types, especially on epithelial cells of the intestine, prostate gland, and the proximal tubules of the kidney (Fleischer, 1994). Localization of ADA and CD26 to the cell surface of lymphocytes increases upon treatment with mitogens, and binding of ADA to CD26 produces a co-stimulatory response in T-cell activation, indicating an extra-enzymatic role of ecto-ADA (Cordero et al., 2001). The increase of CD26 and ecto-ADA at the cell surface also seems to be required to sustain the activation of the T-cell (Franco et al., 1997).

As mentioned previously, extracellular adenosine acts through cell-specific receptors to produce different physiological effects, and therefore its concentration must be tightly regulated (Franco et al., 1997). The amount of extracellular adenosine at any given time is dependent on the activity of adenosine transporter molecules that transport adenosine in and out of the cell, along with the activity of ecto-ADA (Franco et al., 1997). Besides CD26, ecto-ADA can be anchored to the cell membrane by the A_1R adenosine receptor (Figure 1-4). A model of ecto-ADA function that depends on the adenosine concentration has been proposed (Franco et al., 1997). The model suggests that, at low adenosine concentrations, ecto-ADA is available to interact with A_1R and allow high-affinity binding of adenosine, and signal transduction occurs. At these low concentrations, adenosine would not be significantly degraded by ecto-ADA. If extracellular adenosine accumulated, however, the interaction of ecto-ADA and A_1R would be prevented, since the occupation of the ADA active site prevents its binding to A_1R. This prevents high-affinity binding of adenosine, and results in an inefficient signal transduction. Also, the high amounts of adenosine would be degraded by ecto-ADA and thus down-regulate the signal (Franco et al., 1998). Therefore, although there is always a low-affinity binding site for adenosine on A_1R, a high-affinity adenosine binding site and its subsequent signal transduction is only available when ecto-ADA interacts with A_1R to change its conformation (Franco et al., 1998). Overall, the model implies that ecto-ADA is enzymatically inactive when bound to A_1R, but it is active when bound to CD26 (Franco et al., 1997), suggesting that the binding molecules might serve as a regulatory mechanism of ADA activity. Interestingly, in all types of rodent cells studied, CD26 does not interact with ecto-ADA, and significant amounts of A_1Rs are not expressed in hamster cells, suggesting that other receptors are involved in binding ecto-ADA at the surface of rodent cells (Franco et al., 1998).

If ADA1 is a globular cytoplasmic protein, how does it get to the outside of the cell to become ecto-ADA? The answer to this question is not precisely known, but the expression at the cell surface of ecto-ADA is up-regulated by certain cytokines (Cordero et al., 2001). Immune system modulators, including IL-2, IL-12 and IL-4, have been found to play a regulatory role in the translocation of ADA toward the cell surface through a Golgi-independent process that also does not involve CD26 (Cordero et al., 2001). ADA1 might also be transported to the cell surface by a mechanism that does not require a hydrophobic signal sequence, since other proteins have been found to be secreted without a canonical signal at the N-terminus (Muesch et al., 1990).

The third isoform, ADA2, is a 114 kDa dimer that has different kinetic properties and tissue distributions compared with the other two forms, suggesting that it is coded by a separate gene of unknown structure and chromosomal location (Ungerer et al., 1992; Figure 1-4). ADA2 is found in ADA-SCID patients, proving that it results from a separate gene other than ADA1 (reviewed in Hirschhorn and Ratech, 1980). ADA2 is present as a minor fraction of total ADA activity in normal human tissues (Hirschhorn and Ratech, 1980), but its activity can be distinguished from ADA1 and ADA1+CP by use of the selective inhibitor EHNA (Ungerer et al., 1992). Its cellular source is not known and its physiological role is poorly understood. ADA2 may be produced by monocytes, since it makes up 18% of the total ADA activity in these cells (Ungerer et al., 1992), and ADA2 represents the major ADA activity in human serum (Hirschhorn and Ratech, 1980), which suggests it is secreted. This form of ADA has been found in various tissues including liver and spleen, although its proportion of the total activity in these tissues is lower (12% and 2%, respectively) compared with the other forms (24% and 86% for ADA1; 59% and 10% for ADA1+CP) (Van der Weyden and Kelley, 1976). These differences in activity might be due to the fact that ADA2 has a lower affinity (i.e. a higher K_m value) for adenosine than ADA1 does (Hirschhorn and Ratech, 1980; Andreasyan et al., 2005).

Serum ADA activity is increased in patients with various diseases, such as hepatitis, mononucleosis, tuberculosis, pneumonia, and rheumatoid arthritis (Ungerer et al., 1992). The increased ADA activity has been attributed to ADA2, and this up-regulated form of ADA2 purified from human tuberculosis pleural fluid shows the same molecular and kinetic properties as ADA2 from human blood serum (Andreasyan et al., 2005). Why is ADA2 up-regulated when the body is infected? While ADA1 is ubiquitous, ADA2 has been mainly found in monocytes and their descendants (i.e. macrophages), suggesting a role for ADA2 in the immune system. The presence of ADA2 in monocytes/macrophages is puzzling, however, since the intracellular conditions are not optimal for ADA2 function (Gakis, 1996). ADA2 has an optimum pH of 6.5 and a weak affinity for 2-deoxyadenosine compared with adenosine (2-deoxyadenosine/adenosine deamination ratio of 0.25), whereas within monocytes/macrophages the pH is higher than the optimum, making the deamination of 2-deoxyadenosine inefficient in these cells (Gakis et al., 1998). In contrast, ADA1 has an optimal pH of 7-7.5 and a 2-deoxyadenosine/adenosine deamination ratio of 0.75. Since adenosine and 2-deoxyadenosine are toxic to macrophages, they are kept at very low concentrations in these cells by the action of ADA1 (Gakis et al., 1998). It has been suggested that the presence of both ADA1 and ADA2 in monocytes/macrophages is an evolved mechanism for adenosine and 2-deoxyadenosine homeostasis. When an infection occurs, ADA2 levels are increased over ADA1 levels, which allows 2-deoxyadenosine levels to rise and bring about the destruction of nucleic acids in the parasite, thus destroying it (Gakis, 1996). Monocytes/macrophages in the activated state can tolerate high levels of 2-deoxyadenosine, thus the ADA1-ADA2 homeostatic system may act as a tool to produce a "weapon" (2-deoxyadenosine) in monocytes/macrophages against offending parasites (Gakis, 1996).

Protein structure and active site residues of ADA1

The three-dimensional structure of mouse ADA has been resolved, and displays an α/β-barrel structure with eight central β strands and eight peripheral α helices (Wilson et al., 1991). There are also five additional helices that form a lid over an oblong-shaped deep active site. A zinc atom was discovered within the active site upon crystallization of the protein, which is thought to bind an activating water molecule to initiate the reaction (Wilson et al., 1991; Wang and Quiocho, 1998). Several important residues have been identified as contributing to the ADA activity in the mouse protein studied (Figure 1-5). His15, His17, His214 and Asp295 are thought to be important for zinc binding, while His17, Gly184, Glu217, His 238, and Asp296 are important for forming or removing a hydrogen bond within the active site (Wilson et al., 1991; Chang et al., 1991; Sideraki et al., 1996; Mohamedali et al., 1996). Ser265 may form a salt link with His238, while Leu58, Phe61, Leu62, and Phe65 may all be involved in forming a "cap" over the active site pocket (Wilson et al., 1991).

The α/β-barrel structure of ADA is also shared with adenine deaminase (ADE), which catalyzes a mechanistically similar deamination (Figure 1-6), converting adenine to hypoxanthine (Ribard et al., 2003). Another similar reaction, the formation of IMP from AMP, is performed by AMP deaminase (AMPD), and all three enzymes have been suspected to be related through evolution (Becerra and Lazcano, 1998). There have been three AMPD isoforms found, all equally related to ADA, and each encoded by a different gene and expressed in a particular tissue: AMPD1 in muscle, AMPD2 in liver, and AMPD3 in erythrocytes (Gross, 1994). ADE and AMPD also share some of the ADA active site residues (Wilson et al., 1991; Ribard et al., 2003). Only prokaryotic and fungal ADEs have been discovered, presumably because higher organisms do not require the ADE function (Ribard et al., 2003). Also, many bacterial species have apparently lost the ADE gene, since it is not present in species such as *H. influenzae* or *M. pneumoniae* (Becerra and Lazcano, 1998), indicting that either its function is not as essential as it once was, or that other genes are compensating.

ADGF family members

Many proteins with sequence similarity to both *H. sapiens* CECR1 and ADA have been described in the literature, previous to and throughout the course of this graduate project, and as such are described briefly here.

Sarcophaga peregrina IDGF (Insect-Derived Growth Factor), which was later renamed to *S. peregrina* ADGF-A (Zurovec et al., 2002), was initially purified from the culture medium of an embryonic cell line where it acted as a secreted growth factor in an autocrine manner (Homma et al., 1996). The protein was shown to be present as a homodimer, with a subunit molecular mass of 52 kDa. It was expressed in unfertilized eggs, embryos, and first instar larvae, which suggested that it might be important for early developmental stages (Homma et al., 1996; Homma et al., 2001). The growth factor activity of *S. peregrina* ADGF-A was shown to be dependent on its ADA activity using an ADA inhibitor, 2'-deoxycoformycin (DCF), and by mutating two different residues required for ADA activity (Homma et al., 2001). The growth rate of the embryonic cells could also be increased by the addition of ADA derived from calf spleen, but not the addition of the ADA reaction product inosine (Homma et al., 2001), suggesting that it is the breakdown of adenosine that aids growth. Localization of radioiodinated ADGF-A to the cell surface indicated that it might bind a specific molecule in order to exert its growth factor activity (Homma et al., 2001).

Mollusk-Derived Growth Factor (MDGF, formerly called AGSA) was isolated from *Aplysia californica* atrial glands as a 57 kDa glycoprotein (Sossin et al., 1989). The protein was specifically localized to specialized secretory vesicles in the atrial gland, suggesting that it is secreted (Sossin et al., 1989). The atrial gland of *A. californica* is a secretory organ located in the wall of the large hermaphroditic duct (Painter et al., 1985). Differential expression of MDGF in the developing CNS but not in the adult CNS suggested that it may play a role in neuronal growth. MDGF might also have a role in injury repair, since it was upregulated in damaged adult CNS tissue (Akalal and Nagle, 2001). Atrial gland affinity-purified MDGF has been shown to have ADA activity and to stimulate embryonic insect cell proliferation *in vitro* (Akalal et al., 2003). ADA activity from the developing CNS was not tested.

Lutzomyia longipalpis salivary gland ADA (LuloADA) was isolated from a salivary gland cDNA library that was being sequenced in its entirety (Charlab et al., 2000). Salivary gland extracts showed ADA activity more prominently in the lumen than intracellularly, and the activity was significantly reduced following a blood meal. Both of these results indicate that the salivary protein with ADA activity is secreted (Charlab et al., 2000). Recombinant LuloADA was expressed in insect cells and the concentrated culture medium displayed ADA activity, again suggesting that the protein is secreted (Charlab et al., 2001).

A Drosophila homologue, called Male-specific IDGF (MSI), which was later renamed to ADGF-A2 (Zurovec et al., 2002), was shown to be expressed exclusively in mature spermatocytes in the adult testes, suggesting that it plays a role in spermatogenesis (Matsushita et al., 2000). When ADGF-A2 was recombinantly expressed in a Drosophila cell line, it was recovered exclusively in the membrane fraction, and localized to the cell surface, confirming the predicted transmembrane domain near the N-terminus. ADGF-A2 has been the only member of the ADGF family thus far to be predicted as a membrane protein. Recombinant ADGF-A2 exhibited growth factor activity when the cells were confluent, as indicated by radiolabeled thymidine incorporation, suggesting that cell contact was required for the effect (Matsushita et al., 2000). The ADA activity of this protein was not determined.

The cDNAs of the Tsetse Salivary Growth Factors (TSGF–1 and –2) (Li and Aksoy, 2000) were isolated from the salivary glands of *Glossina morsitans*. RT-PCR and Western analyses showed that while both transcripts were also expressed and translated in the midgut, only TSGF-2 was detected in ovary and testes tissues (Li and Aksoy, 2000). Both proteins were also detected in the saliva, suggesting they are secreted. Although the proteins were not tested directly, adenosine deaminase activity was discovered in the salivary gland extracts of *G. morsitans*, and the authors attributed the activity to the two proteins (Li and Aksoy, 2000).

ADA activity has also been discovered in both *Culex quinquefasciatus* and *Aedes aegypti* salivary gland extracts (Ribeiro et al., 2001). A cDNA was identified from a *C. quinquefasciatus* salivary gland library that, when translated, showed sequence similarity to ADA, a predicted signal sequence, and a relative molecular weight of 55 kDa (Ribeiro et al., 2001). Salivary gland fractions containing proteins of this approximate molecular weight showed ADA activity, but the cDNA was not directly tested (Ribeiro et al., 2001). *A. aegypti* was also shown to possess salivary gland ADA activity that was secreted while feeding, whereas secretion could not be detected in *C. quinquefasciatus* (Ribeiro et al., 2001).

Altogether, these proteins make up a growing family of novel growth factors, with sequence similarity to adenosine deaminase (ADA). This gene family has been named ADGF, for Adenosine Deaminase-related Growth Factor. The gene products in this family may exert their growth-factor function through the catalytic conversion of adenosine to inosine. The fact that enzymatic activity is required to stimulate cell proliferation is a novel property of growth factors. Since some ADGF proteins have been shown to exhibit ADA activity, and the cytological location of ADA2 has not yet been found, there may be a connection between these two protein groups that is waiting to be uncovered.

Phylogenetic inference

Phylogenetic analysis can be used to address a number of biological questions, and can provide details on how one group of genes is related to another, as in the case of the ADGF and ADA gene products that show sequence similarity. There are three basic methods of phylogenetic analysis, each with their own advantages and disadvantages, and each method has associated computer software to carry out the analysis, again with specific advantages and disadvantages. The following is an overview, based on a review by M. Holder and P. Lewis (Holder and Lewis, 2003), of the three methods: Maximum parsimony, Distance methods, and Maximum likelihood. Maximum parsimony (MP) is a simple and fast phylogenetic method that counts differences in characters (DNA bases or amino acid residues, in the case of molecular data), ~~and~~ constructs a tree that minimizes the number of overall changes between taxa. Distance methods, such as Neighbour-joining and Minimum evolution, work by converting all sequences into a distance matrix based on the number of changes between each pair of sequences within the set. The tree is constructed using these pair-wise distances. This method is also fast, but the observed differences may not reflect the true amount of evolutionary distance between two sequences, and therefore this method does not perform well on highly diverged sequences. Maximum likelihood (ML) uses one of a variety of models of evolution to search for a single tree that maximizes the likelihood of the data, given that tree. This probability value is called the likelihood function. Once a single tree is found, a bootstrap analysis is performed to measure the amount of support for each branch. Each bootstrap replicate involves the random resampling of characters from within the alignment, and the search for an optimal tree with the resampled data. After many replicates (typically 100-1000), a consensus tree is constructed with the topology found most often within the replicates, and the bootstrap values placed on the tree indicate the number of replicates where that clade was represented. Since bootstrap proportions are conservative measures of support, a value of 70% might indicate strong support for a group (Holder and Lewis, 2003).

ML analysis takes into account a model of evolution, and is therefore a better measure of character evolution than other methods, but is extremely slow, and can only be used on relatively small data sets (Huelsenbeck et al., 2001). Bayesian inference is a more recent phylogenetic method based on ML except that instead of searching for a single optimal tree, it samples trees according to their posterior probability (Huelsenbeck et al., 2001). The posterior probability, or the probability of the tree given the data, can be interpreted as the probability that the tree is correct, and is equal to the product of the likelihood function (probability of the data given the tree) times the prior probability (Huelsenbeck et al., 2001). The prior probability is set out in the model of substitution chosen for the analysis. Although it is impossible to calculate the posterior probability analytically, it can be approximated using the Markov chain Monte Carlo (MCMC) algorithm (thought of as a chain). Each link in an MCMC chain involves two steps, with the goal to find a tree with a better associated probability than the current tree: 1) A new tree is proposed by rearranging the topology and/or branch lengths of the current tree, and 2) the new tree is either accepted or rejected based on its probability. If the new tree is accepted, then it becomes the template for the next link in the chain. When the chain has been run for an adequate length, new trees will not usually be accepted since the current tree has the same probability. At this point, the chain is said to have reached stationarity, and the trees being sampled from then on will fluctuate around a specific probability value (Huelsenbeck et al., 2001).

The computer program MrBayes performs Bayesian analysis by implementing Metropolis coupling to improve the MCMC sampling (Ronquist and Huelsenbeck, 2003). Each analysis consists of four chains that are started from a random tree and run simultaneously. One chain is "cold," in that it is constrained to pick new trees that are very close to the current tree, while three of the chains are "heated," which are free to make more drastic changes to the current tree in an attempt to find a much better new tree (Ronquist and Huelsenbeck, 2003). If a heated chain has a higher probability than the cold chain, the states are swapped such that the better (heated) chain becomes the new cold chain. Whether or not to change states between the cold chain and one of the heated chains is dealt with in each link of the chain. Ideally, the cold chain should visit each of the four locations with equal frequency, since all four chains should eventually converge and fluctuate around an optimum probability value once the "burn-in" period is over. The "burn-in" period includes the collection of tree samples from the beginning of the run that was not optimal (had lower probabilities compared with the end result). Once enough tree samples have been collected, a consensus tree can be constructed after discarding the "burn-in" samples, and the probability of each individual clade found among all the sampled trees is summed (to give the posterior probability value) and indicated on the branches of the consensus tree. Thus, Bayesian analysis is equivalent to performing ML analysis with bootstrap resampling, but it occurs much faster and is therefore useful for inferring large trees (Huelsenbeck et al., 2001). The length of time to run the chain in order to obtain a good approximation of the posterior probability, however, remains the most difficult problem associated with Bayesian analysis (Ronquist and Huelsenbeck, 2003).

Although both MP and likelihood-based methods (ML and Bayesian MCMC) perform well most of the time, MP is strongly biased towards recovering an incorrect tree in situations where highly diverged single taxa are present, since MP tends to erroneously group these taxa together. This phenomenon is called "long branch attraction" (Holder and Lewis, 2003). Likelihood-based techniques, however, are only guaranteed to recover the true tree when the correct model is used (Huelsenbeck et al., 2001). Since each method has specific advantages and disadvantages, it is advisable to use each method separately to infer the phylogeny of the data set, and compare the outcomes. Also, using amino acid data instead of nucleotide sequences substantially increases the accuracy of all methods, due to convergence being less likely with 20 than with 4 possible states (Kolaczkowski and Thornton, 2004).

Intron positions and the early/late debate

Phylogenetic analysis can be improved by considering additional molecular markers, such as intron positions (Krauss et al., 2005). Introns are intervening sequences of noncoding DNA in a gene, and are often considerably longer than exons. Splicing machinery in the cell removes the introns from mRNA before it is translated into protein (Griffiths et al., 1996). These spliceosomal introns are widespread in eukaryotic genomes, but absent from prokaryotes (Venkatesh et al., 1999). There are two opposing theories about the origin of spliceosomal introns. The "introns-early" point of view suggests that introns were present in primitive coding sequences (before the divergence of prokaryotes and eukaryotes) and that they play a role in exon shuffling, an important mechanism in the evolution of proteins (reviewed in Mattick, 1994). Proponents for this theory suggest that present-day prokaryotes have streamlined their genomes by getting rid of introns, as a consequence of intense competitive pressures in the microbial environment (Mattick, 2004). The "introns-late" theory states that introns were added later, in the lineage leading to eukaryotes, and that these introns allowed exon shuffling to bring about variety and complexity in eukaryotic genes compared with prokaryotes (Mattick, 1994).

The main argument in favour of the introns-early hypothesis is the observed correlation of intron position and the boundaries of structural protein domains (Fedorov et al., 2001). But many introns are also found within protein domains, and it has been found that only phase 0 introns (inserted after the third codon position) are correlated with the structure of ancient (highly conserved among all organisms) proteins (de Souza et al., 1998). There are some instances, however, where the intron position and phase is conserved despite low protein sequence similarity (Betts et al., 2001), which adds to the introns-early theory. Also, there are more phase 0 introns in ancient genes than would be expected by chance, and intron positions seem to be located more frequently between units of protein tertiary structure (reviewed in de Souza, 2003). It has been argued by the introns-late side that the introduction of introns allowed the shuffling of exon domains in higher organisms, which accounts for a larger proteome without vastly increasing the genome size (Liu and Grigoriev, 2004). The receptor tyrosine kinase (RTK) gene from an ancient marine sponge contains no introns in the main section of the protein, while all known genes from higher animals contain several introns, thus also supporting the introns-late theory (Gamulin et al., 1997). But almost all of the introns in the integrin-β gene from a coral species were retained in at least one other phylum, suggesting that different introns were lost in the various higher animals, which therefore supports the introns-early hypothesis (Schmitt and Brower, 2001). Some introns in the mosquito triose phosphate isomerase (TPI) gene were thought to be ancient, while others are clearly more recent additions (Tyshenko and Walker, 1997). These authors stated though that the ancient origin of an intron is all but impossible to prove due to the fast rate of intron divergence, intron-sliding, and intron loss. Clearly, there are valid points for both sides of the argument.

The last piece of evidence in the previous paragraph brings up another aspect to the intron debate called intron-sliding, which is the apparent shift of an entire intron by only a few base pairs (Stoltzfus et al., 1997; Rogozin et al., 2000). Spliceosomal introns are not self-splicing and are not known to be mobile, so the loss or gain of an intron in a specific lineage is likely to be a unique event (Venkatesh et al., 1999). Therefore, it might be more parsimonious for the intron to slide rather than be lost and then gained in a nearby location (Schmitt and Brower, 2001). The two intron theories were compared using the aldehyde dehydrogenase (ALDH) genes, and it was found that the data required far too many intron slippage events in order to be consistent with the introns-early model, and it was thus concluded that the ALDH data supported the introns-late theory (Rzhetsky et al., 1997). For the integrin-β gene from the coral species described above to support the introns-early theory, many of the intron positions conserved in other phyla rely on intron-sliding to be considered conserved (Schmitt and Brower, 2001). Although one group concluded that intron-sliding by one base pair might be a real evolutionary phenomenon, they suggest that it would be a relatively rare event occurring in <5% of all introns (Rogozin et al., 2000), so the use of this theory for one or the other side of the debate may be a moot point.

In summary, it seems clear that many introns have indeed been introduced in the lineage leading to eukaryotes, thus validating the introns-late side of the debate. The fact remains, however, that some intron positions of ancient genes have been retained in their descendants, and there is a correlation between phase 0 introns and protein domains (de Souza, 2003). In order to reconcile these two statements, a new "synthetic" theory of intron evolution that incorporates concepts from both theories has been proposed. This new theory suggests that most introns, especially those that are phase 1 or 2, are recent acquisitions of eukaryotic genes, but a subset of the present-day phase 0 introns are candidates to be ancient (de Souza, 2003). Although this new theory still needs to be tested extensively, and does not totally end the debate between early and late, it might be a step in the right direction towards advancement of the field. The study of intron positions can also add to the analysis of the phylogenetic relationships between genes with sequence similarity, such as the ADGF and ADA families, since it represents another aspect of conservation in addition to sequence similarity.

Research objectives

The main goal of this project was to further characterize the *CECR1* gene in order to ascertain its role in the production of the CES phenotype when overexpressed. The specific objectives undertaken in order to accomplish this goal were as follows:

1. Participate in the characterization of the CES critical region, including the determination of the gene structure of the *IL-17R* gene.
2. Identify the 3.5 kb band present on the *CECR1* Northern blot.
3. Create *CECR1* transgenic mice to determine if overexpression of *CECR1* results in features consistent with CES.
4. Design and test antibodies raised against human CECR1 to examine CECR1 expression in transgenic mice and determine the sub-cellular localization of CECR1 to establish if the protein is secreted.
5. Employ RNA *in situ* hybridization of zebrafish and pig embryos to narrow down the spatial and temporal expression pattern of *CECR1*.
6. Using the preceding information as a guide, confirm the developmental profile of *CECR1* expression discovered in model organisms using a limited set of human embryo sections.
7. Verify the existence of the putative antisense transcript to *CECR1* in both pig and human.
8. Characterize the structure and expression patterns of the six Drosophila *CECR1* homologues (*ADGF* genes) in collaboration with Dr. John Locke's lab.
9. Construct the phylogenetic relationship of CECR1 from a variety of organisms, in relation to proteins with significant sequence similarity.

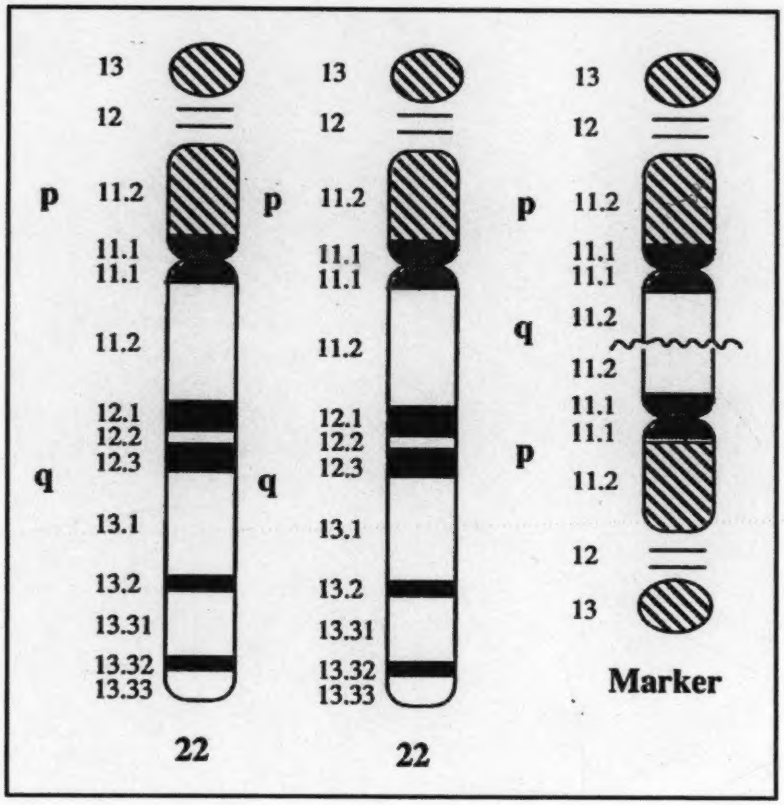
13
12
p
11.2
11.1
11.1
11.2
12.1
12.2
12.3
q
13.1
13.2
13.31
13.32
13.33
22
13
12
p
11.2
11.1
11.1
11.2
12.1
12.2
12.3
q
13.1
13.2
13.31
13.32
13.33
22
13
12
p
11.2
11.1
11.1
q
11.2
11.2
11.1
11.1
p
11.2
12
13
Marker

Figure 1-1. Complement of genetic material from chromosome 22 in a typical CES patient. CES patients normally carry two complete copies of chromosome 22, plus a marker chromosome. The marker chromosome depicted here shows the two extra copies of part of 22q11.2, along with the p arms. This diagram was obtained from Alan Mears' PhD thesis, 1995.

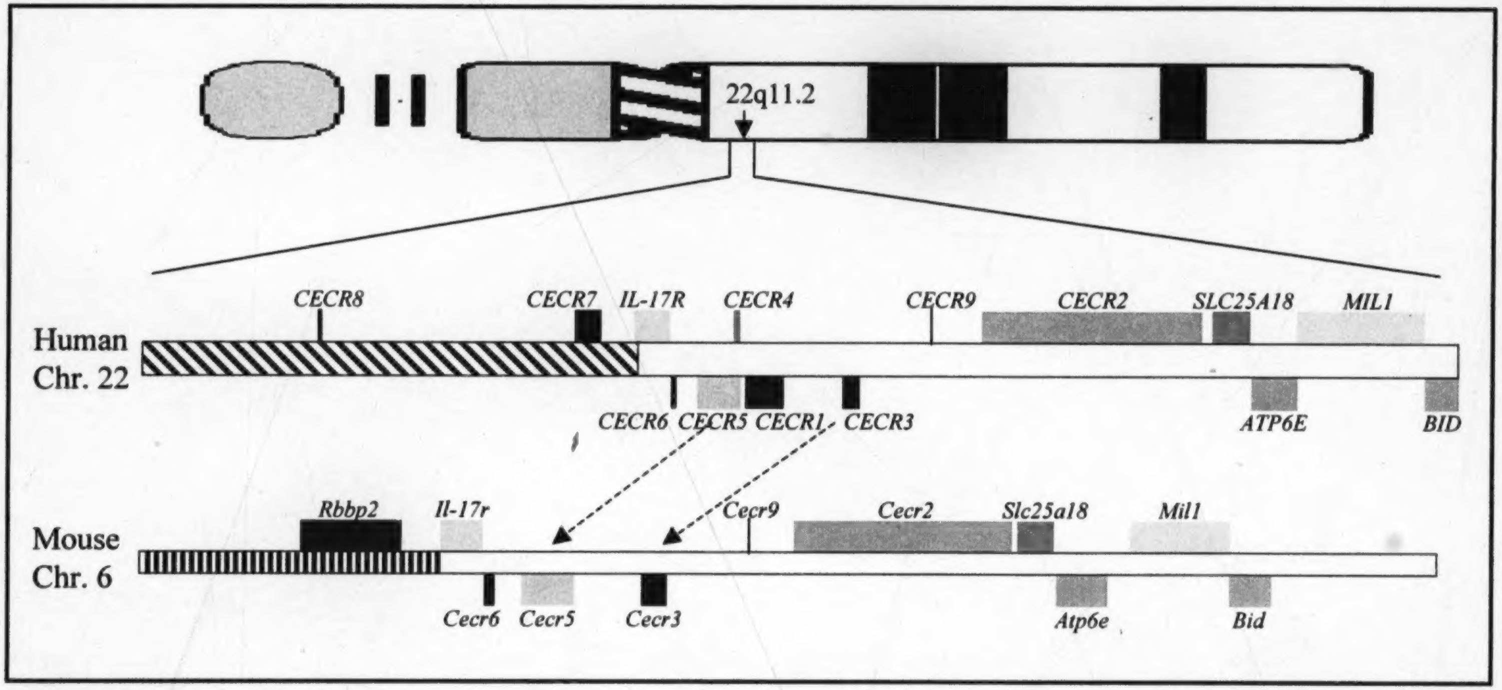
22q11.2
Human
Chr. 22
CECR8
CECR7
IL-17R
CECR4
CECR9
CECR2
SLC25A18
MIL1
CECR6
CECR5
CECR1
CECR3
ATP6E
BID
Mouse
Chr. 6
Rbbp2
Il-17r
Cecr9
Cecr2
Slc25a18
Mil1
Cecr6
Cecr5
Cecr3
Atp6e
Bid

Figure 1-2. Putative genes identified in the CES critical region and syntenic region in mouse. The approximate location of the CES critical region (within 22q11.2) is shown on the chromosome 22 ideogram at the top of the diagram. Coloured boxes represent identified genes, with those transcribed centromere to telomere above the chromosome and those transcribed in the opposite direction below the chromosome. The hatched section represents the pericentromeric region rich in duplications that is not present in the mouse. The banded section represents the portion of mouse chromosome 6 orthologous to human chromosome 12p13. The arrows highlight the missing *CECR1* homologue in mouse. Adapted from Footz, et al., 2001.

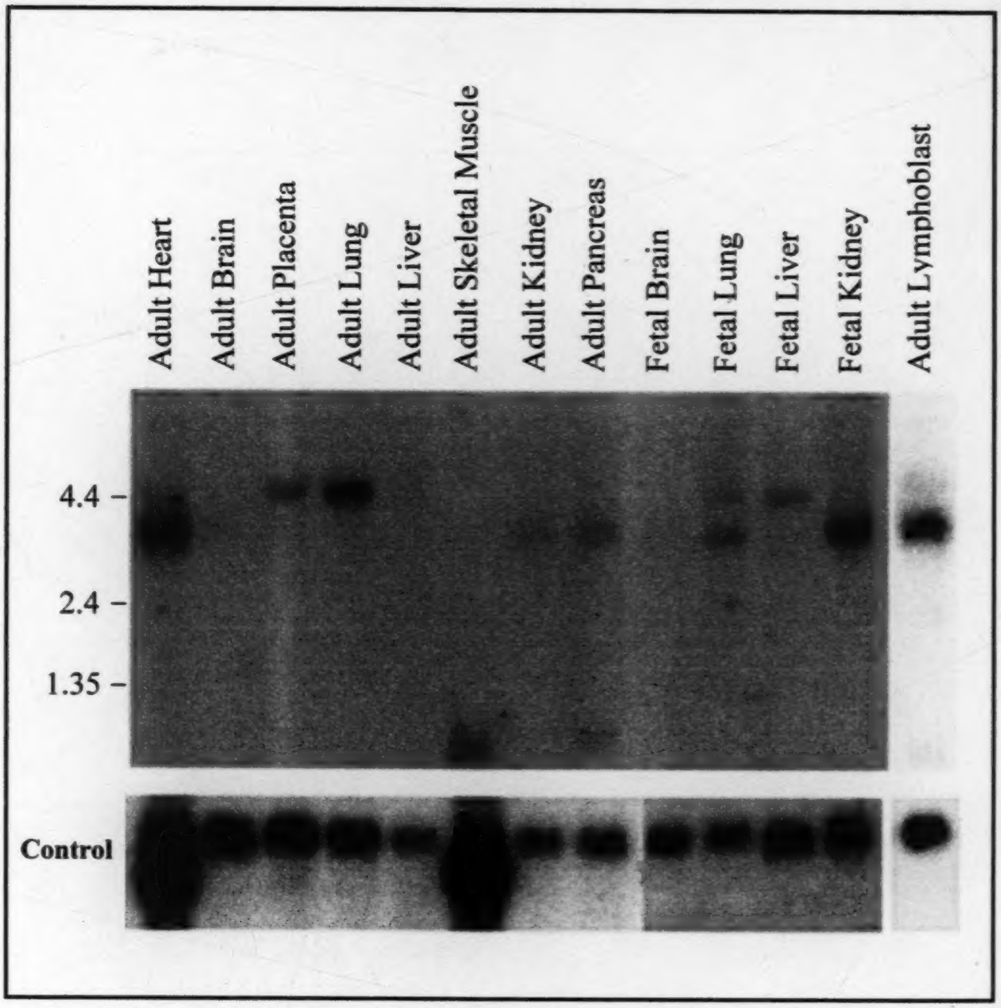
Adult Heart
Adult Brain
Adult Placenta
Adult Lung
Adult Liver
Adult Skeletal Muscle
Adult Kidney
Adult Pancreas
Fetal Brain
Fetal Lung
Fetal Liver
Fetal Kidney
Adult Lymphoblast
4.4 -
2.4 -
1.35 -
Control

Figure 1-3. Northern analysis of human *CECR1*. Adult and fetal Northern blots were hybridized with a probe made from the entire open reading frame of *CECR1*. Transcripts of 4.4 kb and 3.5 kb are visible. The β-actin (adult tissues) or GAPDH (fetal tissues) loading control is shown beneath the blot. Modified from Riazi et al., 2000.

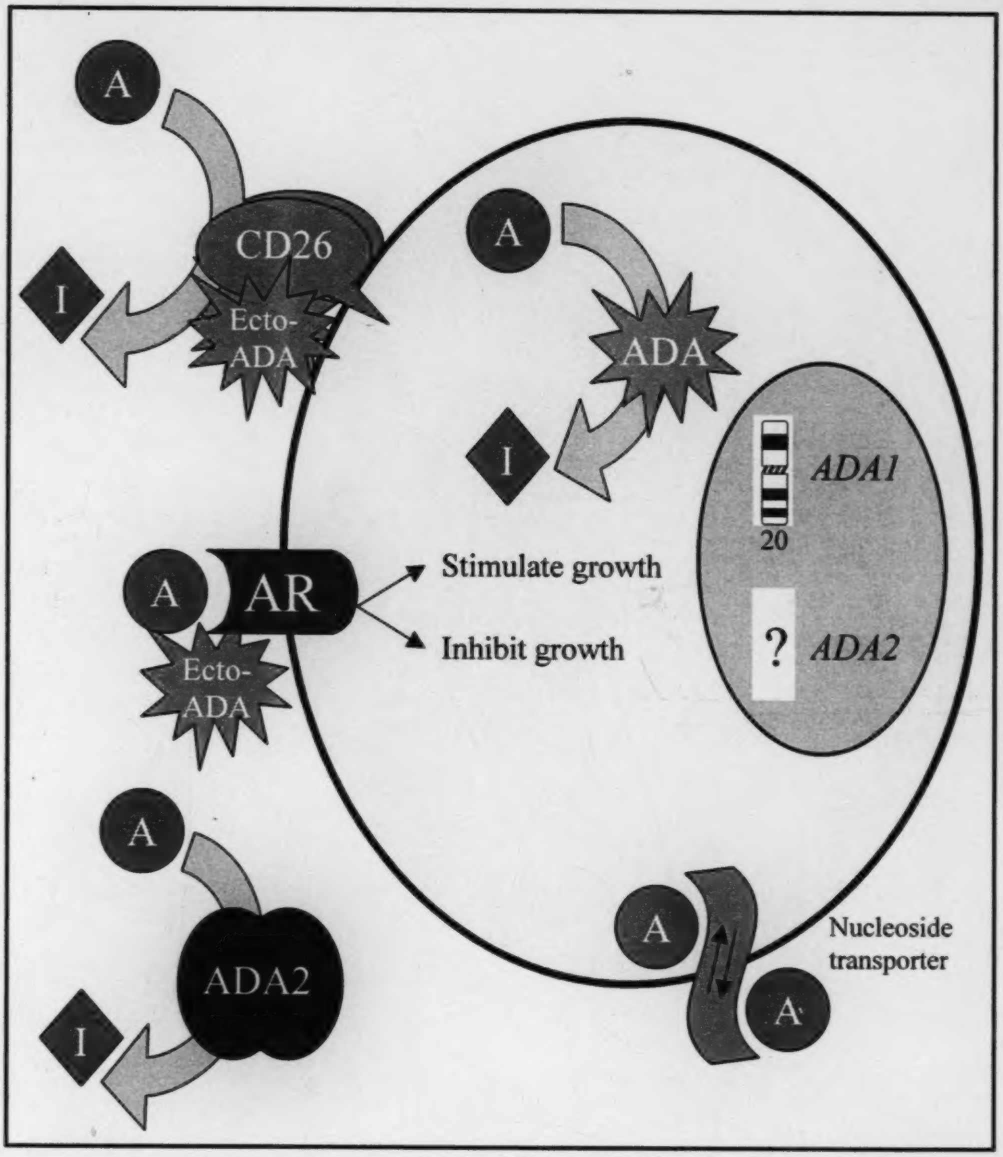
A
CD26
I
Ecto-
ADA
A
ADA
I
ADA1
20
A
AR
Stimulate growth
Inhibit growth
Ecto-
ADA
?
ADA2
A
ADA2
I
A
Nucleoside
transporter
A

Figure 1-4. Schematic depiction of the proteins involved in ADA activity. ADA is a cytosolic protein encoded on chromosome 20 that converts adenosine (A) to inosine (I). Adenosine binds to one of its receptors (AR) to stimulate or inhibit growth. Ecto-ADA combines with CD26 to form the ADA1+CP isoform. Ecto-ADA can bind certain ARs to modulate their affinity for adenosine binding. ADA2 is a secreted dimer, encoded by a gene of unknown chromosomal location.

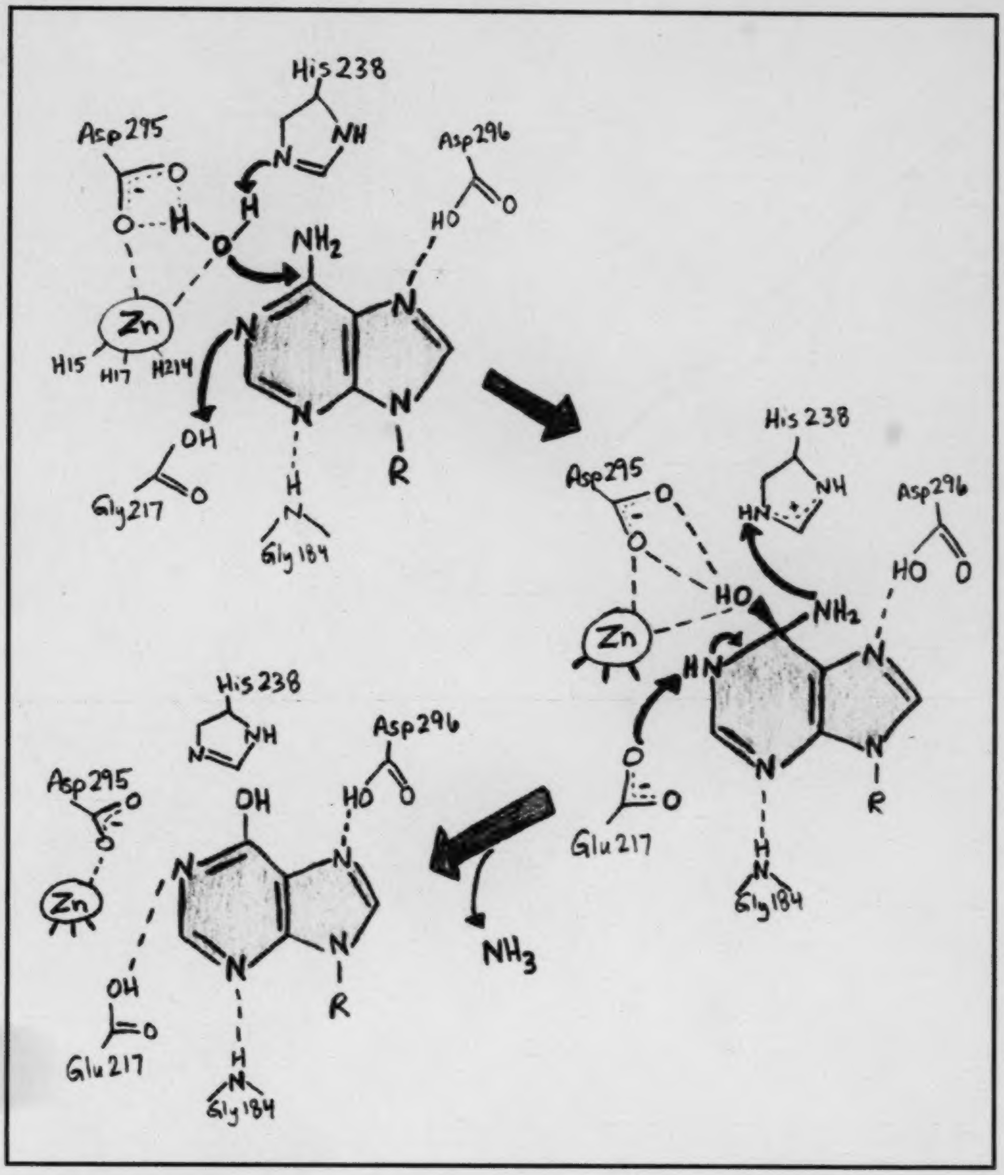
His238
Asp295
Asp296
NH
N
O
H
NH₂
Zn
H15
H17
H214
OH
Gly217
R
Gly184
His238
Asp295
Asp296
HO
NH₂
HN
Glu217
NH₃
His238
Asp295
Asp296
OH
HO
Glu217
Gly184

Figure 1-5. Schematic drawing of the reaction mechanism of adenosine deaminase. Dashed lines indicate non-covalent interactions between neighbouring atoms. Redrawn from Sideraki et al., 1996.

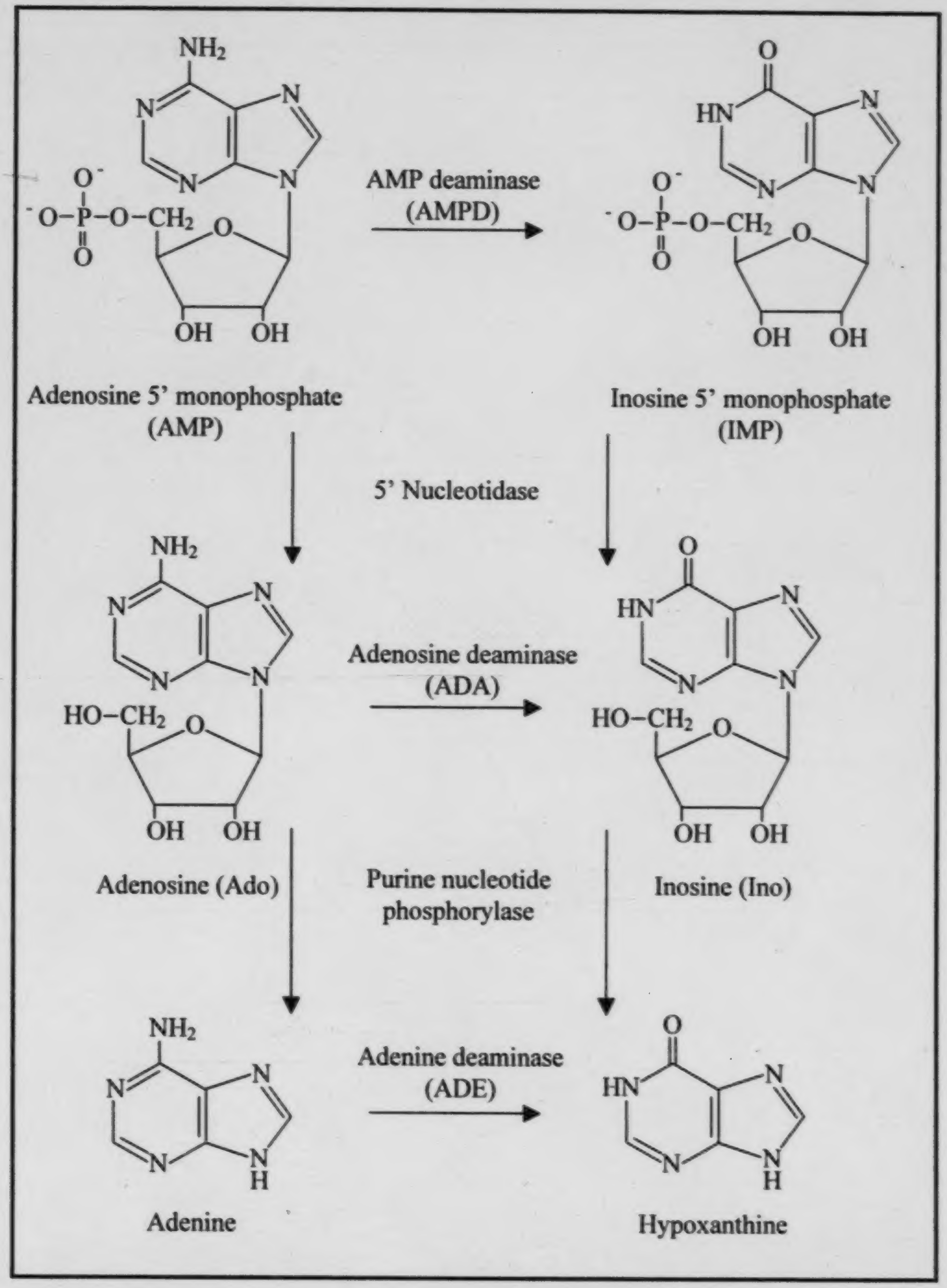
AMP deaminase
(AMPD)
Adenosine 5' monophosphate
(AMP)
Inosine 5' monophosphate
(IMP)
5' Nucleotidase
Adenosine deaminase
(ADA)
Adenosine (Ado)
Purine nucleotide
phosphorylase
Inosine (Ino)
Adenine deaminase
(ADE)
Adenine
Hypoxanthine

Figure 1-6. Reaction mechanisms involved in adenine nucleotide catabolism. This diagram was modified from Becerra & Lazcano, 1998, and M. King's website (http://www.med.unibs.it/~marchesi/nucmetab.html).

Chapter 2: Materials and Methods

Isolation of nucleic acids

Plasmid DNA

Bacterial plasmid DNA was isolated by the traditional alkaline-lysis method (Sambrook and Russell, 2001) from a 5 ml overnight culture usually grown in the presence of 50 μg/ml ampicillin at 37°C in a shaking incubator.

Genomic DNA

Genomic mouse DNA was isolated for genotyping purposes. Briefly, each tail biopsy (~0.5 cm) was digested by 300 μg proteinase K in 350 μl digestion buffer (50 mM Tris pH 8.0, 100 mM EDTA, 100 mM NaCl, 1% SDS) overnight at 56°C. Protein was precipitated by the addition of 125 μl of 5 M NaCl and incubation on ice for 5 minutes. After spinning at 14,000 rpm for 5 minutes to pellet the protein, the supernatant was mixed with 300 μl of isopropanol to precipitate the DNA. Centrifugation for 15 minutes at 14,000 rpm was followed by a 70% ethanol wash, drying, and incubation in 55 μl TE (100 mM Tris, 10 mM EDTA, pH 8.0) at 37°C for 1-2 hours to dissolve the DNA. The genomic DNA sample was then used directly for Southern analysis or PCR.

DNA embedded in agarose

Isolation of bands cut from agarose gels was accomplished with the GeneClean kit (Bio101) as per the manufacturer's instructions. Briefly, the agarose sample was heated in 3 volumes of 6 M NaI at 45-55°C, just until dissolved. An appropriate volume of Glassmilk (usually 10 μl) was added and the sample was mixed gently and placed on ice for 5 minutes to bind the DNA. The glassmilk-DNA was pelleted at 14,000 rpm for 30 seconds and washed three times with 400 μl NEW wash buffer. The DNA was eluted from the glassmilk in 10-15 μl water for 2 minutes at 60°C, followed by centrifugation for 2 minutes at 14,000 rpm. The supernatant, which contained the DNA, was removed to a new tube and stored at –20°C or used directly in downstream procedures.

RNA isolation

Total RNA was isolated from mouse, pig, and human tissues using Trizol reagent (Invitrogen) as per the manufacturer's instructions. Briefly, 0.1 – 0.2 g tissue was homogenized in 1 ml Trizol using an Ultra-Turrax homogenizer, incubated at room temperature for 5 minutes, then clarified at 11,000 rpm for 10 minutes at 4°C. The supernatant was mixed with 200 μl chloroform and incubated at room temperature for 3 minutes, then centrifuged for 15 minutes. The supernatant was mixed with 1 volume (600 μl) isopropanol, incubated for 10 minutes at room temperature, and centrifuged for 15 minutes at 4°C. The pellet was washed with 1 ml ice-cold 75% ethanol in 0.1% DEPC-treated water, centrifuged for 10 minutes at 4°C, dried for about 5 minutes at room temperature, and resuspended in 50-100 μl DEPC-treated water. RNA concentration and integrity was checked using a spectrophotometer at 260 nm and 280 nm.

For the Drosophila Northerns, mRNA was purified using the polyATtract mRNA isolation kit (Promega) and the Oligotex mRNA kit (Qiagen) as per the manufacturer's instructions.

Preparation of BAC clone to create *CECR1* transgenic mice

Human BAC 609c6 was isolated using the QIAGEN plasmid purification kit (QIAGEN), with modifications (Chrast et al., 1999). A large volume (500 ml) of bacteria containing the BAC was grown overnight in LB medium (1% Tryptone, 0.5% Yeast extract, 1% NaCl) plus 10 μg/ml chloramphenicol and centrifuged at 5000 rpm for 10 minutes in a Sorvall GSA rotor to collect the cells. Cells were resuspended in 25 ml QIAGEN Buffer P1 with 100 μg/ml RNase A, then 25 ml QIAGEN Buffer P2 was gently mixed in and incubated for 5 minutes at room temperature. After the addition of 25 ml QIAGEN Buffer P3 and incubation on ice for 20 minutes, the solution was centrifuged at 9000 rpm for 30 minutes at 4°C, then the supernatant was filtered through gauze into a new tube. The DNA was precipitated with 50 ml isopropanol and 5 ml 3 M NaOAc and pelleted at 9000 rpm for 25 minutes at 4°C, then resuspended in 2 ml TE (100 mM Tris, 10 mM EDTA, pH 8.0) before the addition of 10 ml QIAGEN Buffer QBT. The DNA solution was bound onto a QIAGEN tip-500 column that had been previously equilibrated with 10 ml QIAGEN Buffer QBT. The column was then washed twice with

30 ml QIAGEN Buffer QC, and the DNA was eluted into an Oakridge tube with 15 ml QIAGEN Buffer QF that was preheated at 65°C. After the addition of 10.5 ml isopropanol and 1 ml 3 M NaOAc, the mixture was centrifuged at 10,000 rpm for 30 minutes, washed with 1 ml 70% ethanol and centrifuged again at 12,000 rpm for 10 minutes. The DNA pellet was air-dried for 5 minutes and then resuspended in 500 µl injection buffer (10 mM Tris pH 7.5, 0.1 mM EDTA) overnight at room temperature.

The BAC DNA was then purified on a sepharose column using a protocol developed by Angela Johnson (McDermid lab). To prepare the purification column, 25 ml Sepharose CL-4B (Pharmacia Biotech) was equilibrated with an equal volume of TE for 1 hour on a rocker table, allowed to settle, and then new TE was exchanged and rocked overnight. A glass column (25 cm long, 1 cm in diameter) was filled with the sepharose slurry and allowed to settle, after which the column was perfused with injection buffer for 1 hour at a rate of 3-5 drops per minute. To prepare the sample, 90 µl of Blue dye (6% glycerol, 0.2% Bromophenol blue) was added to 450 µl of the BAC DNA. The sample was then loaded on the column, after which 1 ml of injection buffer was allowed to enter the column. The loaded column was then perfused with injection buffer while collecting 500 µl elution samples on ice for a total of 10 fractions.

The samples were run on a pulsed-field gel electrophoresis (PFGE) apparatus to determine which fractions contained the BAC DNA, and to check its integrity. A 0.8% agarose gel was prepared in 0.5X TBE (45 mM Tris, 45 mM Boric acid, 1 mM EDTA) and the samples were loaded alongside the Yeast Chromosome PFGE marker (New England Biolabs). Electrophoresis was carried out at 200 volts for 20 hours and the pulse time varied continuously from 1 to 8 seconds. Fractions 3 and 4 contained the majority of purified BAC DNA obtained from the sepharose column. The BAC DNA from both fractions appeared to be of high quality (low amount of degradation) and was estimated to be approximately 2.5 µg/ml by comparison to a sample of known concentration.

The BAC 609c6 DNA from fraction 3 was used in various concentrations by Dr. Peter Dickie (HSLAS, University of Alberta) for pronuclear injection (Hogan et al., 1994). Briefly, the DNA was injected into the male pronucleus of fertilized eggs harvested from a fertilized FVB/N female. Multiple injected eggs were then implanted into a number of pseudopregnant FVB/N females that were mated to a vasectomized male by Dr. Dickie. Tail biopsies from resultant pups were then tested for the presence of the BAC using Southern analysis to identify founders that would initiate the transgenic line.

DNA probe preparation

DNA probes for Southern or Northern analyses were prepared in the same manner. The DNA fragments were digested with appropriate restriction enzymes or amplified by PCR using specific primers (see Table 2-1) and then purified on a 0.8% low-melt agarose (Sea Plaque, FMC) gel by electrophoresis. The isolated DNA fragments were labeled using either the Strip-EZ DNA or PCR kit (Ambion). For the Strip-EZ DNA procedure, the low-melt agarose plug containing the DNA fragment was boiled for 10 minutes and then 9 μl was added to a 37°C pre-incubated mixture of reagents provided in the kit: 2.5 μl 10X Decamer solution, 5 μl 5X Buffer -dATP/-dCTP, and 2.5 μl 10X modified dCTP. After the addition of 1 μl Klenow (Ambion) and 5 μl (α-P^{32}) dATP (Amersham Biosciences), the mixture was incubated at 37°C for 15 minutes, then passed through a Sephadex G-50 column to remove unincorporated nucleotides. The flow-through containing the double-stranded DNA (dsDNA) probe was then added to the blot for hybridization.

The Strip-EZ PCR kit (Ambion) was used to make a single-stranded (ssDNA) probes. For this procedure, the PCR template was combined with 10 pmol of the appropriate primer (see Table 2-1), 1 μl 10X PCR buffer, 1 μl dNTPs (both provided in the kit), 0.5 μl *Taq* polymerase (Microbiology Department, University of Alberta), and 2.5 μl (α-P^{32}) dATP (Amersham Biosciences) in a 10 μl reaction. Labeling occurred during the PCR reaction, which consisted of 40 cycles of 94° for 20 seconds, T_m for 20 seconds, and 72° for 1 minute. The reaction was then passed through a Sephadex G-50 column and the flow-through was added to the blot for hybridization.

Southern analysis

Genomic or plasmid DNA samples were digested with the appropriate restriction enzyme followed by electrophoresis on a 0.8% agarose (Invitrogen) gel in 1X TAE (40 mM Tris-acetate, 1 mM EDTA, pH 8.4). The DNA within the gel was denatured by washing the gel in denaturing solution (0.5 M NaOH, 1.5 M NaCl) twice for 20 minutes, neutralized (0.5 M Tris pH 7.5, 3 M NaCl) for 30 minutes, rinsed in 10X SSC for 15 minutes, and capillary transferred to a GeneScreen Plus nylon membrane (NEN Life Science Products) in 10X SSC overnight. The membrane was subsequently washed for 10 minutes each in wash #1 (0.4 N NaOH) and wash #2 (0.2 M Tris pH 7.5, 2X SSC), UV cross-linked for 30 seconds, then dried on Whatman paper. Blots were pre-hybridized in "Westneat solution" (5X Denhardt's solution [50X stock: 1% Ficoll 400, 1% polyvinylpyrrolidone, 1% BSA], 264 mM sodium phosphate pH 7.5, 1 mM EDTA, 6% SDS) modified slightly from the published formulation (Westneat et al., 1988) for 1-2 hours at 65°C in a roller-bottle hybridization oven (Tyler Research). The P^{32}-labeled probe was then added and hybridization occurred at 65°C overnight. Cross-species hybridization was performed at lower temperatures (50-55°C). Two low stringency washes (1.5X SSC, 0.2% SDS) at room temperature were followed by one or two high stringency washes (0.2X SSC, 0.2% SDS) at 65°C, depending on the strength of the signal. Rarely, the blot was washed in a higher stringency buffer (0.1X SSC, 0.2% SDS). The membrane was then exposed to Biomax XAR film (Kodak) at –70°C for an appropriate length of time (typically 2-5 days).

Northern analysis

Human adult and fetal multiple tissue Northern blots were obtained from Clontech Laboratories, which contained approximately 2 µg of poly(A)+ RNA per lane. Other Northern blots were prepared by electrophoresis of total RNA or poly(A)+ mRNA on an agarose-formaldehyde gel (1.2% agarose, 1.85% formaldehyde [37% stock], 1X MOPS [20 mM MOPS, 2 mM NaOAc, 1 mM EDTA, pH 7.5]). The RNA samples (~40 µg per lane) were heated at 60°C for 15 minutes in sample buffer (50% formamide, 6.5% formaldehyde, 1X MOPS) to remove secondary structure before loading on the gel along with 1/5th volume of RNA loading dye (50% glycerol, 1 mM EDTA pH 8.0, 0.25% bromophenol blue, 0.25% xylene cyanol). The samples, alongside a 0.24 - 9.5 kb RNA ladder (Invitrogen), were electrophoresed in 1X MOPS at 100 V until the bromophenol blue dye was within an inch from the bottom. The ladder lane was removed and stained in an ethidium bromide solution for 15 minutes, followed by destaining in Milli-Q (ddH_2O) water overnight before photographing. The rest of the gel was rinsed in Milli-Q water for 45 minutes, then in 10X SSC (in 0.1% DEPC-treated water) for 45 minutes, followed by capillary transfer onto a GeneScreen Plus nylon membrane (NEN Life Science Products) in 20X SSC (in DEPC-treated water) overnight. The membrane was subsequently baked at 80°C in a vacuum oven for 2 hours.

The blot was pre-hybridized in Northern hybridization solution (50% formamide, 5X SSPE, 10X Denhardt's solution, 2% SDS, 0.4 mg/ml herring sperm DNA) in a sealed bag for 1 hour at 42°C before hybridization to a P^{32}-labeled probe at 42°C over two nights in a shaking water bath. Two low stringency washes (2X SSC, 0.1% SDS) at room temperature were followed by one or two high stringency washes (0.1X SSC, 0.1% SDS) at 50°C, depending on the strength of the signal. The membrane was then exposed to Biomax XAR film (Kodak) at –70°C for an appropriate length of time (typically 1-2 weeks) dependant on the signal strength.

In order to probe a Northern sequentially, the previous probe was stripped off the blot using the buffers supplied with the labeling kit (Ambion). The blot was stripped for 10 minutes at 68°C in 10 ml 1X DNA Probe Degradation Buffer (in 1X Probe Degradation Dilution Buffer), followed by a 10 minute wash at 68°C in 10 ml 1X Blot Reconstitution Buffer and 0.1% SDS. The blot was then incubated with pre-hybridization buffer before the new probe was added.

RT-PCR

Total RNA was isolated from tissues using the Trizol (Invitrogen) method. Approximately 1 μg of RNA was treated with 1 U DNase1 (Invitrogen) for 15 minutes at room temperature in 1X DNase reaction buffer with 40 U RNaseOUT (Invitrogen) to remove any contaminating DNA. The reaction was stopped by adding 1 μl of 25 mM EDTA and incubating for 10 minutes at 65°C. Using the ThermoScript RT-PCR System (Invitrogen), the RNA solution was mixed with 1X cDNA synthesis buffer, 5 mM dithiothreitol, 1 mM dNTPs, 15 U ThermoScript reverse transcriptase, and 2.5 μM oligo$(dT)_{20}$ reverse primer. The reaction was incubated in a PTC-100 programmable thermal cycler (MJ Research) in the following scheme to make single-stranded DNA:

42°C, 30 minutes, 10 minutes each of 50°C, 53°C, 55°C, 57°C, and 60°C, followed by 5 minutes at 85°C. If a gene-specific primer was used as the reverse template, the five temperatures used in the series ranged from 50°C to 65°C. 2 U RNase H was added and incubated for 20 minutes at 37°C to remove residual RNA. An aliquot of this reaction was subsequently used in a standard PCR reaction. As a negative control, each RNA sample was also carried through this procedure without reverse transcriptase present in the first-strand cDNA synthesis reaction, in order to test for amplification from genomic DNA.

PCR

A standard PCR reaction contained 1X PCR Buffer (25 mM Tris pH 9.0, 50 mM KCl, 1.5 mM $MgCl_2$, 0.02 mg/ml BSA fraction V), 0.2 mM dNTPs, 0.2 μM of each primer (see Table 2-1), an appropriate amount of DNA, and 1 U of *Taq* polymerase (Microbiology Department, University of Alberta). PCR reactions were carried out in a PTC-100 or PTC-200 programmable thermal cycler (MJ Research). The PCR program typically followed the "touchdown" (Don et al., 1991) procedure: 94°C, 2 minutes; (94°C, 30 seconds; ($T_m$initial) – 0.6°C/cycle, 30 seconds; 72°C, 1 minute per kb) for 10 cycles; (94°C, 30 seconds; $T_m$final, 30 seconds; 72°C, 1 minute per kb) for 25-30 cycles; 72°C, 10 minutes; 4°C indefinitely. PCR products were often cloned into the pGEM-T Easy vector (Promega, see Appendix Figure A1) for sequencing purposes.

Sequencing

Sequencing of PCR products and cDNA clones for identification purposes was carried out using the Thermo Sequenase radiolabeled terminator cycle sequencing kit (Amersham Biosciences) and run manually on an 8% polyacrylamide gel. Sequencing of vector-ligated DNA inserts was carried out using dye-labeled M13 forward and reverse primers with the Thermo Sequenase fluorescent labeled primer cycle sequencing kit (Amersham Biosciences) and run on a Li-Cor automated sequencing apparatus. Alternately, sequencing was completed on an ABI 377 automated sequencer (Applied

Biosystems) using vector or clone-specific primers along with the fluorescently labeled DYEnamic ET Terminator Cycle Sequencing kit (Amersham Biosciences).

Screening a zebrafish cDNA library

A 19-25 hpf (hours post fertilization) zebrafish lambda-ZAP cDNA library was obtained from Dave Pilgrim, University of Alberta (originally from Bruce Appel, University of Oregon). *E. coli* XL1-Blue plating cells were grown overnight in LB containing 0.2% maltose and 10 mM $MgSO_4$ at 37°C and then resuspended in 10 mM $MgSO_4$ to an OD_{600} of 0.5. Various dilutions of phage in SM buffer (100 mM NaCl, 8 mM $MgSO_4$-$7H_2O$, 50 mM Tris) were mixed with plating cells and 0.7% top agarose on 10 cm LB plates to titer the library at 1.16×10^9 pfu/ml. The library was then plated out on twenty 15 cm LB plates (~18,000 plaques each) that were grown for approximately 8 – 10 hours at 37°C, then incubated at 4°C for several hours.

Plaques were lifted onto Hybond-N (Amersham Biosciences) nylon membranes by placing the membrane on the plate surface for 1 minute. The membrane was then placed DNA-side up on Whatman paper soaked in denaturing solution (1.5 M NaCl, 0.5 M NaOH) for 3 minutes, neutralization buffer (1.5 M NaCl, 0.5 M Tris pH 7.5) twice for 3 minutes, and then agitated in 2X SSC for 5 minutes. The membranes were then dried on Whatman paper and baked for 2 hours at 80°C in a vacuum oven. Plaque hybridization to a P^{32}-labeled probe was carried out in Hybond hybridization solution (5X SSC, 5X Denhardt's solution, 0.5% SDS) at 65°C overnight. Two low stringency washes (2X SSC, 0.1% SDS) at room temperature were followed by two high stringency washes (1X SSC, 0.1% SDS) at 65°C, and three higher stringency washes (0.1X SSC, 0.1% SDS) at 65°C, 68°C, and 70°C, respectively, since the background signal was very strong. The membranes were then exposed to Biomax XAR film (Kodak) at –70°C for 3-5 days.

One positive plaque was isolated into SM buffer from the original plate, re-plated at a low density on 10 cm LB plates for the secondary screen, and then the plaque-lift and probe hybridization procedures were repeated. A single plaque was isolated into SM buffer and *in vivo* excision was carried out. Briefly, XL1-Blue plating cells were mixed with the positive phage stock and ExAssist Helper phage (Stratagene), incubated at 37°C for 15 minutes, and then grown in LB for 3 hours at 37°C to amplify the phage. After lysing the cells and lambda phage at 65°C for 20 minutes and centrifuging the debris, the phagemid supernatant was incubated in different amounts with SOLR cells for 15 minutes at 37°C, then the mixtures were plated onto LB plates and grown overnight at 37°C. Colonies were isolated and the pBluescript phagemid (Appendix Figure A2) containing the insert was obtained

Western analysis

Antibody production to CECR1

A recombinant CECR1 fusion protein was made for injection into rabbits to produce an anti-CECR1 antibody. A PCR reaction using the *CECR1* cDNA (IMAGE clone 54445; AF190746) and primers HIDExp-F and HIDExp-R (Table 2-1) was carried out to amplify the DNA encoding amino acids 30 to 302 (see also Figure 3-7). The PCR reaction was run using Platinum *Taq* DNA Polymerase High Fidelity (Invitrogen) with 1.5 mM final $MgSO_4$ concentration, and the DNA insert was subsequently subcloned into the *Bam*HI and *Eco*RI sites of the bacterial expression vector pRSET A (Invitrogen; see Appendix Figure A3). Bacteria containing this HIDExp1 construct were grown in SOB medium (2% tryptone, 0.5% yeast extract, 8.6 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$) at 37°C until OD_{600}=0.3, induced with 1 mM IPTG and grown at 37°C for 30 minutes, then infected with 5 pfu/cell M13-T7 phage and grown for 5 hours to express the HIDExp1 fusion protein. The fusion protein was extracted from the cell pellet with Extraction Buffer (6 M Guanidine-HCl, 0.5 M NaCl, 50 mM Tris, pH 7.8) and purified on a Probond Ni-Agarose (Invitrogen) column by first washing with Buffer A (8 M Urea, 0.5 M NaCl, 50 mM Tris, pH 7.8) and Buffer A plus 40 mM Imidazole, and then eluting with Buffer A plus 150 mM Imidazole. The sample was dialyzed overnight in Milli-Q water and then lyophilized overnight in a freeze dryer (Virtis), before resuspending the HIDExp1 recombinant protein in sterile Milli-Q water to a concentration of 2 mg/ml. The fusion protein was separated on an SDS-PAGE gel, and the ~32 kDa size protein band (predicted weight: 36.5 kDa) was cut out for injection into rabbits (0A1, 0A6), along with Freund's adjuvant (Sigma). Rabbit serum was collected at various intervals and tested using Western blots.

Two CECR1 peptides were chosen in order to make rabbit anti-peptide antibodies that would hopefully recognize the native CECR1 protein. Several criteria were followed in the design, especially the avoidance of hydrophobic or oxidation-prone residues, complicated secondary structure, residues that may be post-translationally modified (cysteines, glycosylation, phosphorylation), clusters of residues with bulky side chains, and homopolymers of one amino acid. The two peptides chosen, CECR1-Pep1 (AHPTPRPSEK) and CECR1-Pep2 (GETDWQGTSI), were synthesized and conjugated to both KLH and BSA by the Alberta Peptide Institute (API, University of Alberta). The KLH-conjugated peptides were injected into two rabbits each (Pep1: rabbits 2F6 and 2F2, Pep2: rabbits 2F1 and 2F3). Serum was collected at various intervals and tested using ELISA and Western blots.

Protein sample preparation and quantification

Proteins were extracted from various human and mouse tissues at 4°C using RIPA buffer (50 mM Tris pH 7.5, 150 mM NaCl, 1% Nonidet-P40, 0.5% sodium deoxycholate, 0.1% SDS, plus a protease inhibitor tablet; Boehringer Mannheim) at a concentration of 10^7 cells/ml or 100 mg tissue/ml. The cells were homogenized in the RIPA buffer briefly with an Ultra-Turrax homogenizer, and cell lysates were clarified at 14,000 rpm for 10 minutes at 4°C to remove cellular debris. Alternately, a denaturing protein extraction was performed, by homogenizing the sample at 10^8 cells/ml or 1 g tissue/ml in denaturing lysis buffer (2% SDS, 50 mM Tris pH 7.5), followed by boiling for 10 minutes and sonication for 5 bursts at maximum intensity. The mixture was clarified at 10,000 rpm for 10 minutes, then the supernatant was diluted 10-fold with Triton-X dilution buffer (2% Triton-X 100, 50 mM Tris pH 7.5) and incubated on ice for 10 minutes before clarifying once more. All protein supernatants were kept either at –20°C for short-term or –70°C for long-term storage until required.

Total protein concentrations were determined using the DC Protein Assay kit (Bio-Rad) with bovine serum albumin (BSA) as a standard control. The protocol was scaled down so that only 20 µl of protein sample, 100 µl of Reagent A, and 800 µl of Reagent B were used per assay. After incubating the mixture for 15 minutes, the absorbance was measured at A_{750} on a spectrophotometer.

Western gel electrophoresis and transfer

Samples for the SDS-PAGE gel were boiled with 1/6th volume of 6X loading buffer (0.35 M Tris-Cl pH 6.8, 10% SDS, 30% glycerol, 0.6 M DTT, 0.06% bromophenol blue) and separated on a 7.5% SDS-PAGE separating gel with a 4% stacking layer. The gel was electrophoresed in running buffer (0.025 M Tris, 0.192 M glycine, 0.1% SDS) at 25 mA per 1.0 mm gel using the Hoefer SE 600 apparatus (Amersham Biosciences) until the bromophenol blue dye reached the end of the gel. The proteins were then transferred to a Hybond-ECL nitrocellulose membrane (Amersham Biosciences) in Towben buffer (25 mM Tris, 192 mM glycine, 0.1% SDS, 20% methanol) at 75 mA overnight in the cold room. The BenchMark Protein Ladder (Invitrogen) and total proteins loaded in other lanes were visualized by incubating the membrane in 1X Ponceau S solution (20X stock: 2% Ponceau S in 30% trichloroacetic acid and 30% sulfosalicylic acid; Sigma) for 15-30 minutes, followed by destaining in 5% acetic acid. Gels that were not transferred were rinsed three times with water, stained with GelCode Blue Stain Reagent (Pierce) for 1 hour, and then destained overnight in water.

Western detection of proteins

All steps of the following detection protocol were carried out at room temperature on a rocker or shaker table. The protein membrane was blocked with Blotto (5% skim milk powder (Carnation) in PBS/T [Western formulation: 80 mM Na_2HPO_4 (anhydrous), 20 mM NaH_2PO_4, 100 mM NaCl, 0.1% Tween-20, pH 7.5]) for 1.5 hours, washed twice for 5 minutes in PBS/T, then incubated with a CECR1 antibody in Blotto at the appropriate dilution for 1.5 hours. After washing the excess primary antibody away with four 5 minutes washes in PBS/T, the blots were incubated with a 1:5000 dilution of Goat F(ab_2) anti-rabbit-HRP antibody (CALTAG Laboratories) in Blotto for 45 minutes to an hour. Four 5 minutes washes in PBS/T were again done to remove excess secondary antibody. Detection was carried out using the ECL Western Blotting Analysis system (Amersham Biosciences) after which the blot was exposed to Biomax XAR film

(Kodak), typically for 5 minutes and then a second exposure for between 20 minutes to overnight.

Competition assay

To remove CECR1-specific antibodies from rabbit serum before using the serum in the Western detection protocol, the serum was pre-treated with recombinant HIDExp1 protein as the competing antigen. Various amounts of the HIDExp1 protein were incubated with 1 µl of serum antibodies in a final volume of 200 µl PBS (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 2 mM KH_2PO_4) for 2 hours at room temperature, with occasional mixing, then overnight at 4°C. The mixture was centrifuged at 14,000 rpm for 15 minutes, and the supernatant was added to 3 ml Blotto before using it as the primary antibody solution at a 1:3000 final dilution in the Western detection protocol.

ELISA

The peptide antibodies were tested for recognition of the peptide antigen using the following ELISA protocol. The wells of an Immulon 2H microtiter plate (Dynex Technologies) were each coated with 200 ng BSA-conjugated peptide (either BSA-Pep1 or BSA-Pep2) by incubating 100 µl/well of 2 µg/ml peptide in 0.1 M Na_2CO_3 pH 9.6, for 2 hours at 37°C. The solution was flicked out and the wells were rinsed 3 times with PBS/T (ELISA formulation: 20 mM Na_2HPO_4, 1.5 mM KH_2PO_4, 140 mM NaCl, 3 mM KCl, 0.05% Tween-20, pH 7.4). Excess binding sites were blocked with 350 µl/well Blotto (5% skim milk powder (Carnation) in PBS/T) overnight at 4°C. After rinsing 3 times with PBS/T, various dilutions of rabbit anti-CECR1 serum (Pep1: rabbits 2F6 and 2F2, Pep2: rabbits 2F1 and 2F3) in Blotto were added and incubated for 1.5 hours at 37°C. The wells were rinsed 3 times, and AP-conjugated goat anti-rabbit IgG (Bio-Rad; 100 µl/well, diluted 500- or 1000-fold in Blotto) was incubated in the wells for 1 hour at 37°C, followed by 3 more rinses. Finally, the detection substrate (100 µl/well; one Sigma 104 Phosphatase tablet dissolved in 5 ml 0.1 M Diethanolamine, Sigma) was incubated in the dark at room temperature until sufficient color had developed (about 25 minutes). The reaction was stopped by adding 100 µl/well 0.4 M NaOH, and the absorbency at 405 nm was read on a Thermomax microtiter plate reader.

Protein A pull-down experiment

Preparation of Protein A beads

Protein A sepharose CL-4B beads (Sigma) were reconstituted from powder by swelling in Buffer A (20 mM NaH_2PO_4, 150 mM NaCl) for 30 minutes at room temperature, followed by two 10 minute washes in Buffer A, with centrifugation between steps to remove the supernatant. Note that all centrifugation steps for the following protocol were 10,000g for 30 seconds (or 3000g for 5 minutes). The protein A beads were then stored at 4°C in a 1:1 ratio with Buffer A plus 0.1% sodium azide until required. After pre-washing the prepared protein A beads in PBS, approximately 2 mg of each antibody was bound to every 1 ml of protein A beads in 10 volumes of PBS for 1 hour, rocking at room temperature. The antibody-bound beads were washed twice in 10 volumes of 0.2 M sodium borate (pH 9.0) before incubating in this solution with 20 mM dimethyl pimelimidate (Sigma) for 30 minutes to couple the antibodies to the beads. The coupling reaction was stopped by rinsing the beads in 10 volumes of 0.2 M ethanolamine (pH 8.0) and then incubating the beads in this solution for 2 hours. The antibody-coupled beads were rinsed in PBS and then washed twice in 10 volumes of 1 M glycine (pH 3.0) for 5 minutes each to remove uncoupled antibodies. Finally, the beads were washed three times with 10 volumes of PBS for 5 minutes each, and then stored in PBS in a 1:1 ratio at 4°C. The success of the coupling reaction was tested by running before- and after-coupling samples on an SDS-PAGE gel.

Large-scale protein A precipitation

While usually only 1 ml of protein extract is used for this type of experiment, this protocol describes a "large-scale" method of protein A precipitation used to obtain a sample for mass spectrometry analysis, such that 10 ml of protein lysate was used for each experiment. Note that all centrifugation steps for this protocol were at 3000g for 4 minutes. Each protein lysate was pre-cleared with 200 μl of protein A beads only (not coupled to antibodies), for 1 hour rocking at 4°C. The beads were precipitated and the supernatant was then incubated with 200 μl of protein A-preimmune beads (protein A beads coupled to preimmune antibodies) for 1.5 hours, rocking at 4°C. The beads were again centrifuged and 200 µl of protein A-immune beads were added to the supernatant and incubated for 2 hours at 4°C. After each step, a protein lysate (supernatant) sample was taken for SDS-PAGE and Western analysis to observe if the immune-coupled beads had removed the antigen. The beads from each step were washed three times with 10 ml cold PBS, rocking at 4°C for 5 minutes each, followed by boiling in 1X Laemmli buffer without DTT (2% SDS, 10% glycerol, 60 mM Tris pH 6.8, 0.02% Bromophenol blue) for 5 minutes. After centrifugation, the elutant was removed from the beads, boiled with 100 mM DTT for 2 minutes, and then analyzed on an SDS-PAGE gel stained with GelCode Blue Stain Reagent (Pierce).

Cell culture

A CES patient fibroblast cell line, JGe, was grown in RPMI medium (Invitrogen) with 10% FBS (fetal bovine serum, Invitrogen), 1% L-Glutamine (Invitrogen) and 1% Penicillin-Streptomycin (Invitrogen) with 5% CO_2 at 37°C in 10 cm plates. To split the cells once they were confluent, the cells were washed in medium containing no FBS, then incubated with 2 ml 0.05% Trypsin (Invitrogen) until cells began lifting off the plate. The trypsin was inactivated by adding RPMI medium containing FBS, and the cells were suspended in this mixture before aliquoting to fresh 10 cm plates containing 10 ml medium. When required, cells were frozen down in cryovials by collecting trypsinized cells into a 50 ml tube, centrifuging at 1000 rpm for 10 minutes, followed by resuspension of the cell pellet in 2 ml RPMI/FBS medium containing 10% DMSO for every original plate. The cryovials were filled with 1 ml resuspended cell mixture each and then gradually cooled in the –70°C freezer before transferring to liquid nitrogen for long-term storage.

In situ hybridization

Embryo collection and storage

Pig embryos ranging from 20 to 40 days post fertilization (dpf) were obtained from the University of Alberta farm through collaboration with Dr. George Foxcroft. Embryos were collected in 4% paraformaldehyde in PBS and fixed overnight at 4°C. The larger embryos (28-40 dpf) were injected in several places (head, back, and rear) with fixative to ensure penetration of all tissues before incubation overnight. After fixing, the embryos were washed in PBS, dehydrated in a methanol gradient (5 minutes each of 25%, 50%, 75%, and 100% twice), and then stored in 100% methanol at –20°C until required. For the slide hybridization procedure, embryos were embedded in paraffin wax by first washing three times in ethanol and twice with toluene for 15 minute each, followed by a graded series (30%, 60%, 100%, 100%) of paraffin in toluene at 60°C in a vacuum oven for one hour each. A final 100% paraffin wash was performed overnight, before embedding the embryos into paraffin blocks. Embedded embryos were sectioned at 7 μm using a Reichert-Jung microtome, and mounted on poly-L-lysine coated slides (Electron Microscopy Sciences) by floating the sections in a 42°C water bath and transferring the section onto the slide underneath. The slides were dried at 37°C for 2-3 days and then stored at room temperature until required.

Human fetal kidney and liver tissues were obtained from Dr. Stephen Bamforth, University of Alberta. Tissues were fixed, embedded and sectioned as described for the pig embryos. A limited set of human embryo sections mounted on slides was obtained from Dr. Michel Vekemans, Hopital Necker Enfants malades in Paris, France.

Preparation and testing of digoxigenin-labeled RNA probes

The corresponding DNA fragment used to make each RNA probe was first amplified by PCR using specific primers (see Table 2-1). Each PCR fragment was cloned into the *Eco*RV site of the pBluescript SK- (Stratagene) or pZErO-2 (Invitrogen) vector (see Appendix Figures A2 and A4) or into pGEM-T Easy (Figure A1) in both directions such that both sense and antisense probes could be made separately using the T7 promoter. Alternately, restriction digests were utilized to generate the insert in both orientations. Each fragment-containing plasmid was digested on the opposite side of the insert from the T7 promoter, using *Eco*RI (or *Sal*I for pGEM-T Easy), and electrophoresed on an agarose gel. The linearized plasmid was isolated by the "Freeze and Squeeze" method. Briefly, two freeze-thaw cycles of the agarose slice were followed by collection of the DNA by centrifugation at 14000 rpm for 10 minutes through glass wool covering a needle hole in a 1.5 ml eppendorf tube into another 1.5 ml collection tube. Phenol-chloroform extraction and ethanol precipitation were then carried out as described for plasmid isolation (Sambrook and Russell, 2001) and the DNA pellet was dissolved in an appropriate amount of water.

The RNA probes for *in situ* hybridization were made with digoxigenin (DIG)-labeled UTP (Roche) in an *in vitro* transcription reaction using T7 RNA polymerase (Invitrogen) according to the manufacturer's instructions. The RNA probe was ethanol precipitated and dissolved in DEPC-treated water, and then tested for integrity and size on a 0.8% agarose gel. The probe was also tested for incorporation efficiency of the DIG-labeled UTP by carrying out a mock detection reaction. A serial dilution of the probe from 10^0 to 10^{-4} was spotted on a small piece of GeneScreen Plus (NEN Life Science Products) membrane. The membrane was rinsed in Buffer 1 (10 mM Tris pH 7.5, 150 mM NaCl) for 1 minute and then blocked in 1% Blocking reagent in Buffer 1 (stock: 10% Blocking reagent (Boehringer-Mannheim) in Maleic Acid Buffer [100 mM maleic acid, 150 mM NaCl, pH 7.5]) for 30 minutes at room temperature. The membrane was then washed twice in Buffer 1 for 10 minutes each, followed by incubation with the antibody solution (1:2000 anti-DIG antibody conjugated to alkaline phosphatase (Roche) and 1% blocking reagent (Boehringer-Mannheim) in Buffer 1) for 30 minutes at room temperature. After washing with two changes of Buffer 1 for 15 minutes each, the membrane was equilibrated in Buffer 2 (100 mM Tris pH 9.5, 100 mM NaCl, 50 mM $MgCl_2$) for 5 minutes. The antibody was detected using 0.45% NBT (nitro-blue tetrazolium chloride; Boehringer-Mannheim) and 0.35% BCIP (5-bromo-4-chloro-3-indolyl phosphate; Invitrogen) in Buffer 2 until a purple coloured stain was observed (usually about 15 minutes). The intensity of colour on the dots indicated the relative success of digoxigenin labeling of the probe.

Whole mount in situ *hybridization*

This protocol was modified slightly from a published protocol (Wilkinson and Nieto, 1993) obtained from Dr. Rachel Wevrick's Lab, University of Alberta. All embryos were washed at room temperature on a table shaker for 5 minutes, unless otherwise noted. All solutions used up to and including the hybridization were made using 0.1% DEPC-treated water, and all solutions were filtered with 0.45 μm filters

(Millipore) to prevent dirt particles from sticking to the embryos. The pig embryos were rehydrated in a methanol gradient in PBS/T (75%, 50%, 25%, PBS/T twice; PBS/T *in situ* formulation: 137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 2 mM KH_2PO_4, 1% Tween-20), followed by bleaching in 6% H_2O_2 in PBS/T for 1 hour. After washing three times in PBS/T, the embryos were permeabilized in 20 μg/ml proteinase K in PBS/T for 25 minutes. Embryos were then incubated for 10 minutes with freshly made 2 mg/ml glycine in PBS/T, washed twice in PBS/T, and post-fixed for 20 minutes in 4% paraformaldehyde/0.2% glutaraldehyde in PBS/T. After washing four times with PBS/T, the embryos were incubated in two changes of pre-hybridization solution (50% formamide, 5X SSC (pH 4.5), 1% SDS, 50 μg/ml yeast RNA (Roche), 50 μg/ml heparin (Fisher)) for 1 hour each, rotating in eppendorf tubes within a hybridization oven at 68°C. The hybridization solution (pre-hybridization solution plus 2 μl/ml DIG-labeled RNA probe) was denatured at 80°C for 5 minutes immediately before adding to the embryos for incubation at 68°C overnight.

Excess probe was removed by washing three times in Solution 1 (50% formamide, 5X SSC pH 4.5, 1% SDS) followed by three washes in Solution 2 (50% formamide, 2X SSC pH 4.5) for 30 minutes each, shaking at 68°C. The embryos were washed three times in TBS/T (25 mM Tris pH 7.5, 137 mM NaCl, 2.7 mM KCl, 1% Tween-20), and then blocked with 10% heat-inactivated sheep serum (Sigma) in TBS/T for 2.5 hours. During this step the anti-DIG antibody conjugated to alkaline phosphatase (Roche) was presorbed with pig acetone powder. Briefly, a small amount of acetone powder was heated in TBS/T with 2 mM levamisole (Sigma) at 68°C for 30 minutes, centrifuged at 10,000 rpm for 10 minutes, and then the pellet was incubated with 1:1000 anti-DIG antibody in TBS/T with 1% sheep serum and 2% blocking reagent (Boehringer-Mannheim) at 4°C for 1 hour. After centrifugation at 10,000 rpm for 10 minutes, the supernatant was diluted 2-fold with TBS/T (containing 1% sheep serum and 2 mM levamisole) for use in the procedure. The acetone powder was made previously by crushing an embryo in PBS, incubation in 4 ml cold acetone for 30 minutes, centrifugation for 5 minutes at 10,000 rpm, resuspension of the pellet in 1 ml ice-cold acetone for 10 minutes, centrifugation again, and then air drying the pellet before pulverizing into powder.

After blocking, the embryos were incubated with the 1:2000 presorbed anti-DIG antibody solution on a shaker table at 4°C overnight. Unbound antibody was washed away using TBS/T containing 2 mM levamisole, first for three times of 5 minutes followed by five times of 1.5 hours each, then overnight at 4°C. In order to detect the signal, embryos were first washed three times in AP developing solution (100 mM Tris pH 9.5, 100 mM NaCl, 50 mM $MgCl_2$, 2 mM levamisole, 0.1% Tween-20), followed by incubation with 0.5% NBT (Boehringer-Mannheim) and 0.375% BCIP (Invitrogen) in AP developing solution in the dark for 1 hour or more until a purple coloured stain was observed. To stop the reaction, the embryos were rinsed twice in PBS/T containing 20mM EDTA, followed by many washes (once per hour) in this solution. The embryos were then post-fixed with 4% paraformaldehyde in PBS/T for 1 hour, washed twice with PBS/T, dehydrated and rehydrated in a methanol in PBS/T gradient (25%, 50%, 75%, 100%, 100%, 75%, 50%, 25%, and PBS/T twice) to clarify the signal, and stored at 4°C. Digital images of the embryos were captured on a Nikon digital camera attached to a dissecting scope.

Slide in situ *hybridization*

This *in situ* hybridization protocol was developed by Song Hu in the McDermid lab, based on a journal article (Braissant et al., 1998). Slides were washed and incubated at room temperature in a coplin jar on a table shaker, except where indicated. All solutions used up to and including the hybridization were made using 0.1% DEPC-treated water. Paraffin sections were de-waxed in two changes of toluene for 5 minutes each, followed by re-hydration in a graded ethanol series (100%, 95%, 70%, and water) for 2 minutes each, and two washes of PBS for 5 minutes each. Sections were post-fixed in 4% formaldehyde in PBS for 10 minutes, and then treated with active DEPC water in PBS twice for 15 minutes each. In order to permeabilize the tissue, 20 μg/ml proteinase K in TE (20 mM Tris pH 7.5, 5 mM EDTA) was incubated with the slides for 20 seconds at 37°C. Slides were equilibrated in 5X SSC twice for 10 minutes each before pre-hybridization in hybridization solution (50% formamide, 5X SSC, 500 μg/ml salmon sperm DNA) at 60-65°C in an InSlide Out hybridization oven (Boekel) for 2 hours. During the last few minutes of this incubation, the DIG-labeled RNA probe was prepared in hybridization solution (0.25 µl of probe in 120 µl hybridization solution for one slide) and denatured at 80°C for 5 minutes. The pre-hybridization solution was decanted and replaced with the denatured probe solution, and the slides were incubated overnight at the hybridization temperature (typically 60-65°C).

Excess probe was washed off in three successively stringent washes (2X SSC at room temperature for 30 minutes, and 2X SSC followed by 0.1X SSC at the hybridization temperature for 1 hour each). The slides were equilibrated in Buffer 1 (10 mM Tris pH 7.5, 150 mM NaCl) twice for 5 minutes each before incubation in the antibody solution (1:2000 anti-DIG antibody conjugated to alkaline phosphatase (Roche) and 1% blocking reagent (Boehringer-Mannheim) in Buffer 1) for 2 hours at room temperature. Excess antibody was removed with two washes in Buffer 1 for 15 minutes each, followed by equilibration in Buffer 2 (100 mM Tris pH 9.5, 100 mM NaCl, 50 mM $MgCl_2$) for 5 minutes. Detection of the antibody was achieved through colour development using 0.45% NBT (Boehringer-Mannheim) and 0.35% BCIP (Invitrogen) in Buffer 2 until a purple coloured stain was observed. Typically, slides were incubated for 2 hours at 37°C to initiate the process, and then transferred to either room temperature or 4°C overnight, depending on the progress of colour development. To stop the colour reaction, the slides were incubated in TE twice for 5 minutes each. The slides were counter-stained with 0.5% Methyl Green in 0.1 M NaOAc pH 4.0 for 2 minutes, rinsed three times in water, and then dehydrated in a graduated ethanol series (35%, 50%, 75%, 95%, 100%) and two washes of toluene for 2 minutes each. The sections were mounted in DPX Mountant (BDH Chemicals) with a cover slip and dried overnight before taking pictures with a dissecting or compound microscope coupled to a Nikon digital camera.

Sequence analysis and computer software

Gene discovery, prediction and annotation tools

GeneTool v1.0 or v2.0 (BioTools) was used to analyze and compile sequence chromatographs generated by the ABI or LICOR automated sequencers, compare gene sequences, and predict PCR primers in some cases. Various forms (blastn, tblastn, bl2seq) of the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) were accessed at the National Center for Biotechnology Information (NCBI;

http://www.ncbi.nlm.nih.gov/BLAST) to search the different NCBI genome databases for DNA and protein similarity using either the default parameters, or without "filtering" of low complexity sequence. For the proteins involved in the phylogenetic analysis in particular, the tblastn algorithm was used with one of the subfamily protein sequences to search the GenBank (Benson et al., 2004) non-redundant (nr), EST, or species-specific (both finished and incomplete) genomic databases for similar gene products.

Exon prediction was accomplished with GRAIL2, which is available through the BCM Search Launcher site (http://searchlauncher.bcm.tmc.edu/). GENSCAN (Burge and Karlin, 1997), accessed at http://bioweb.pasteur.fr/seqanal/interfaces/genscan.html, was used for predicting entire cDNA sequences from genomic data. The AntiHunter (Lavorgna et al., 2004b) web program (http://bio.ifom-firc.it/ANTIHUNTER/) was used to search for and predict antisense transcripts in the *CECR1* genomic region. Human genomic interspersed repeats were identified and masked by the RepeatMasker program (A.F.A. Smit & P. Green, unpublished data; http://www.repeatmasker.org).

The ExPASy Proteomics Server (http://kr.expasy.org) suite of programs (Gasteiger et al., 2003) were used for translation (Translate tool), molecular weight prediction (Compute pI/Mw), N-glycosylation prediction (NetNGlyc Server), signal sequence prediction using SignalP (Bendtsen et al., 2004), and cellular localization using TargetP (Emanuelsson et al., 2000). Secondary structure of proteins was predicted using PepTool v1.0 (BioTools) for peptide selection.

Multiple alignment and phylogenetic analysis

Alignment of DNA or protein sequences was carried out using ClustalW (http://www.ebi.ac.uk/clustalw) (Thompson et al., 1994) using the output format "aln wo/numbers." Alternately, protein sequences were aligned using MUSCLE (http://www.drive5.com/muscle)(Edgar, 2004) using the FASTA or ALN output formats. MUSCLE has been shown to produce better alignments at a faster rate than ClustalW (Edgar, 2004). Identical/similar amino acids within an alignment were shaded using BOXSHADE (www.ch.embnet.org/software/BOX_form.html) (K. Hofmann & M.D. Baron, unpublished) using the input sequence format "ALN" and the "RTF new" output format. MUSCLE alignments were checked by eye in MacClade 4.03 (Maddison and

Maddison, 1989), where manual editing and removal of regions with large gaps (stretches of sequence without counterparts in other species) was accomplished.

Maximum Parsimony (MP) analysis was performed using PAUP* version 4.0b8 (Swofford, 2001). A simple amino acid substitution matrix was employed, which counts the minimum number of nucleotide substitutions required to convert one amino acid to another (Figure 2-1), based on the PROTPARS model described in Joe Felsenstein's PHYLIP Manual (http://evolution.genetics.washington.edu/phylip/doc/protpars.html) and formulated by Dr. Warren Gallin, University of Alberta. A heuristic search was performed, using the default parameters, except that 100 search replicates were performed, each started by random stepwise addition of taxa, before branch swapping using tree-bisection-reconnection (TBR). A bootstrap analysis of 100 replicates, each starting from 20 random stepwise additions of taxa, was performed in order to obtain support values for placement on the most parsimonious tree. Due to the large number of taxa, the bootstrap analysis took approximately three weeks.

Bayesian inference was performed using MrBayes v3.0 (Ronquist and Huelsenbeck, 2003) with the Jones model (Jones et al., 1992) of amino acid substitution provided in the package. This substitution model is best for the comparison of sequences that are highly diverged, which was the case here. Prior probabilities for all trees were equal. The Markov chain Monte Carlo (MCMC) sampling was performed with one cold and three heated chains and was usually started from a random tree. Each search was run for 150,000 generations (initial analyses) or 550,000 generations (final analyses), with trees being sampled every 100 generations, and the first 500 tree samples were discarded as "burn-in." Phylogenetic trees were viewed and printed using TreeView 1.6.6 (Page, 1996).

Mapping of intron positions

For all ingroup sequences (predicted or confirmed) discovered, the cDNA sequence was compared against genomic sequence, where available, to determine the locations of exon/intron boundaries within the coding sequence. Introns located outside of the open reading frame (ORF) were not considered. The intron positions were placed on the alignment of ingroup proteins and a matrix was built based on the presence/absence of a given intron location in each protein sequence. At each of the 52 locations found, the presence of an intron was coded as a character type of "1" for each taxa, while absence of an intron at that location was coded as "0." Taxa with no genomic sequence to compare with were coded as "?" for all intron positions. Introns were only considered homologous if they were identical in both location and phase. The ancestral state of each intron position was reconstructed on the rooted Bayesian topology using the "trace character" feature of MacClade (Maddison and Maddison, 1989).

Table 2-1. Primer sequences used in the study

Species_Gene	Primer name	5' - 3' sequence	Design[1]
Hs_IL-17R	IL-F1	GCGCTGGGCGAAATAGCGTC	SM
	IL-F2	CATGGCGTCTCCTGACCTCCTT	SM
	IL-R1	TGGGAGCGGGCTGTGTGGAT	SM
	IL-R2	CCAGTGTCAGACGGCACCTCA	SM
	IL-R3	GGGCCATACACCATCTGGGACA	SM
	IL-R4	GGGTAAGCAAGGACCAGTGTGAGA	SM
	IL-R5	CCGGGTGACTGCCTGCTTTCA	SM
	IL-R6	CGGGAGGCAAGGTCTGAGAGT	SM
Hs_CECR1	HIDGF-5	TCTCCATCTGAGCCCTTTCC (forward)	AJ
	HID-F1	TCCATCTGAGCCCTTTCCTA	SM
	HID-F2	TCATCGCAGATTCCATCCGA	SM
	HID-F3	ATGAAATCAATGGCCTCTGT	SM
	HID-F4	AGAGAAGTCAAGTGTTTA	CJ
	HID-F5	GCTGCTGCCGGTGTATGA	LB
	HID-F6	AAAAGGACATCCCCATAG	LB
	HID-F8	TACCCTGTTGGAGAGTGAGA	SM
	HID-R1	CTCCATACAGAGGCCATTGA	SM
	HID-R2	TGTGAGCTCTCCAAGTGCAT	SM
	HID-R3	TTGAGTACTAACTAGCTTTG	SM
	HID-R6	CCCTTTTGGCATCATCCT	CJ
	HID-R7	TTCCTGGTAAGTCTTCAC	LB
	HID-R8	GCACCTGGTTAGAGATGG	LB
	HIDExp-F	CGCGGATCCATAGATGAAACACGGGCGCATCT	SM
	HIDExp-R	CGGAATTCCCTATTATGCGATGACAGCCACATCTTTG	SM
	HID-I8-A	TGGGCTCCTCCTCTTCCTG	SM
	HID-I8-B	GCAAGGCTCTAATGTCCTCT	SM
	HID-I8-C	TGGCTGTTTAGTCCTTGCTG	SM
	CECR1-Var2-F	TCTGAGCGTCCTAATAGCCA	SM
	CECR1-Var2-R	CAGAGCCCAGTTCCATTCCA	SM
Dm_ADGF-A	GH082-F	CGGCCATTATCAACCTGACCTA	SM
	GH082-R	GCGCACAGTATCCATAGGAGTA	SM
	GH082-R2	GGTAACCAGGCCTTCGTCAA	SM
	GH082-R3	CGCTTCTCAACGTCTTGTAG	SM
	Dros75A-F2	CTTGCGTCTGCTGAAGAAGTTG	SM
	75A5'-F1	TGAAGATCGCGGCGAGGAAGT	SM
	75A5'-F2	AGGAAGTGCTCGGAATCTGA	SM
	75A5'-F3	GTCATATTGTGTGGTCTACG	SM
Dm_ADGF-A2	MSI-F	GCACTCCACAGAATGCAATG	SM
	MSI-F2	GGTCTGCGTTAACTTGGAAC	SM
	MSI-R	CGCGGTAGATCAAATCGATG	SM

	MSI-R2	CCGCCGTTTCCAAAATCGTA	SM
	MSI-R3	CGCCACCATATTTACTGGAC	SM
	MSI-R4	CTCGGACGTAAGATTGGACA	SM
Dm_ADGF-B	75A-L-F	GGACACAGTGGCCATTTACA	SM
	75A-L-R	ATGGGAGATCCGAACCAGTT	SM
	75L-R2	CCGGTAGGTCAGACTTATGA	SM
	75L-R3	CCGATCGTTCAGTCGTATCT	SM
Dm_ADGF-C	LP055-F	CCCGCCGAAATTATACTTGACAT	SM
	LP055-F2	TCGGTGCTGCATGTGCACA	SM
	LP055-R	GGAACTGGGGACTATTCAATCTT	SM
	LP055-R3	TACCTCATGTTCGACCCTCTC	SM
	LP055-R4	GTAGAGCTCCTCCAGCATCT	SM
	Dros87F1-F	CGTCCGTCGTTCACGTCAGT	SM
Dm_ADGF-D	GH122-F	TGCCGAGGATCGGGGAAAGTAC	SM
	GH122-R	GATGCGCGATTCCGCTTGGCAT	SM
Dm_ADGF-E	50A-F	CCAAGCGGGAAACTGTGCAA	SM
	50A-R	GTAGCCATGACCGATCCTCT	SM
Dm_RPS3a	RP5F-1	GTCGTCAACGTGATTCGACCTTTTCCG	LP
	RP3R-1	AATTTAAACAGCTTCCTGTACTGGG	LP
Dm_"ADA"	85C-F	GGGCATCAAAAGCCTATGTGA	SM
	85C-R	TACCGCTAAGATCGATGCCTA	SM
Ss_CECR1	PID-F1	CCAAAGACGTGTCCCTCATC	PB
	PID-F2	TATGAGGCCTTCATGGGTCT	KK
	PID-F3	TCATCTCCGTTTCTTCATC	JS
	PID-F4	CAAAGTGAGCAGAACAGCA	JS
	PID-F5	TATGTATTTGTCCAGGTTTC	JS
	PID-R1	ATCGTAGGACAGGCCTTTGG	PB
	PID-R2	GTGTAGGAGTGGATGAGACA	KK
	PID-R3	AACCTGGACAAATACATAGT	JS
	MHC2b-F1	CAAGCTACTGAGGCAATAAG	SM
	MHC2b-R1	GAACCTCCCGACTCTTGAC	SM
	MHC2b-R2	TGTGCATTTCTTTGGTCAC	SM
Dr_CECR1-1	ZID-F1	CCCCTGTGGTTAAAGAGATG	SM
	ZID-F2	GGTTTCAGTGGATTGGCTGGTG	SM
	ZID-F3	GGGCTCGGGTGATTTTCACTG	SM
	ZID-R1	GCACACTTACAACAAGACAAG	SM
Tr_CECR1-2	Fugu2-F1	AGCCGCAGGAATCTGGATTC	SM
	Fugu2-F2	AGGGACCTGCTGATGCGAGA	SM
	Fugu2-R1	CAGCCGATACTGCCTGCTTC	SM
Xl_CECR1	XenoCECR1-F1	GGGAGATGTGACTGAGTTTG	SM
	XenoCECR1-F2	GGGAGTCAGGTTCTTGTTTG	SM
	XenoCECR1-F3	ACGGTGAAGGGAGCAGAGTT	SM
	XenoCECR1-R1	CTGGGATTCTATTTGGAGTAC	SM
	XenoCECR1-R2	TCCCTGCCAATTTGTCTCTC	SM
	XenoCECR1-R3	ACGCCTCCATCAGCTTCATC	SM

Various vectors	T7	GTAATACGACTCACTATAGGGC	
	T3	AATTAACCCTCACTAAAGGG	
	PM001	CGTTAGAACGCGGCTACAAT	
	SP6	ATTTAGGTGACACTATAGAATACT	

[1] Primers were designed by: AJ, Angela Johnson; CJ, Cheryl Johnson; JS, Jon Staav; KK, Katie Kessler; LP, Lynn Podemski; PB, Polly Brinkman-Mills; SM, Stephanie Maier.

Sequences highlighted in grey depict the restriction enzyme site

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
[A]	.	2	1	1	2	1	2	2	2	2	2	2	1	2	2	1	1	1	2	2
[C]	2	.	2	2	1	1	2	2	2	2	2	2	2	2	1	1	2	2	1	1
[D]	1	2	.	1	2	1	1	2	2	2	2	1	2	2	2	2	2	1	2	1
[E]	1	2	1	.	2	1	2	2	1	2	2	2	2	1	2	2	2	1	2	2
[F]	2	1	2	2	.	2	2	1	2	1	2	2	2	2	2	1	2	1	2	1
[G]	1	1	1	1	2	.	2	2	2	2	2	2	2	2	1	1	2	1	1	2
[H]	2	2	1	2	2	2	.	2	2	1	2	1	1	1	1	2	2	2	2	1
[I]	2	2	2	2	1	2	2	.	1	1	1	1	2	2	1	1	1	1	2	2
[K]	2	2	2	1	2	2	2	1	.	2	1	1	2	1	1	2	1	2	2	2
[L]	2	2	2	2	1	2	1	1	2	.	1	2	1	1	1	1	2	1	1	2
[M]	2	2	2	2	2	2	2	1	1	1	.	2	2	2	1	2	1	1	2	3
[N]	2	2	1	2	2	2	1	1	1	2	2	.	2	2	2	1	1	2	2	1
[P]	1	2	2	2	2	2	1	2	2	1	2	2	.	1	1	1	1	2	2	2
[Q]	2	2	2	1	2	2	1	2	1	1	2	2	1	.	1	2	2	2	2	2
[R]	2	1	2	2	2	1	1	1	1	1	1	2	1	1	.	1	1	2	1	2
[S]	1	1	2	2	1	1	2	1	2	1	2	1	1	2	1	.	1	2	1	1
[T]	1	2	2	2	2	2	2	1	1	2	1	1	1	2	1	1	.	2	2	2
[V]	1	2	1	1	1	1	2	1	2	1	1	2	2	2	2	2	2	.	2	2
[W]	2	1	2	2	2	1	2	2	2	1	2	2	2	2	1	1	2	2	.	2
[Y]	2	1	1	2	1	2	1	2	2	2	3	1	2	2	2	1	2	2	2	.

Figure 2-1. Amino acid substitution matrix used for maximum parsimony analysis. The letters across the top and down the side of the matrix represent the 20 amino acids. Numbers within the matrix represent the minimum number of nucleotide substitutions required to convert one amino acid to another. This matrix was formulated by Warren Gallin, University of Alberta.

Chapter 3: Results

Defining the proximal CES critical region: Genomic annotation of *IL-17R* and *CECR1*

The CES critical region had been cloned into a contig of BACs/PACs (Johnson et al., 1999) which were being sequenced by Bruce Roe at the University of Oklahoma, at the commencement of this project in the McDermid lab. Computer aided sequence analysis completed by various lab members had identified 14 putative human genes in and around the CES critical region (Footz et al., 2001), and further analysis of many of these genes had begun. This thesis describes the characterization of two genes in the proximal CES critical region, *IL-17R* and *CECR1*, of which the latter became the main focus.

Analysis of the 3' end of IL-17R

The data in this section was published as part of a description of genes in the CES critical region: Footz, T.K., Brinkman-Mills, P., Banting, G.S., **Maier, S.A.**, Riazi, M.A., Bridgland, L., Hu, S., Birren, B., Minoshima, S., Shimizu, N., Pan, H., Nguyen, T., Fang, F., Fu, Y., Ray, L., Wu, H., Shaull, S., Phan, S., Yao, Z., Chen, F., Huan, A., Hu, P., Wang, Q., Loh, P., Qi, S., Roe, B.A. and McDermid, H.E. (2001). Analysis of the cat eye syndrome critical region in humans and the region of conserved synteny in mice: A search for candidate genes at or near the human chromosome 22 pericentromere. *Genome Research* 11: 1053-1070.

IL-17R is a T-cell derived cytokine receptor gene that was previously localized to the 22q11.2 region using a radiation hybrid panel (Yao et al., 1997). This published sequence, however, did not have a polyadenylation signal or otherwise define the 3' end of the gene (Figure 3-1A). Northern analysis with a 697 bp PCR probe made with primers IL-F1 and IL-R1 (Table 2-1) showed eight different sized transcripts (1.05, 1.45, 1.9, 2.6, 5.0, 6.3, 8.8, and 10.5 kb) from various human adult tissues (Figure 3-2A), indicating possible alternative splicing or polyadenylation. There were two EST clusters found distal to the coding sequence that were proposed to be associated with the *IL-17R* cDNA sequence (Figure 3-1A). Northern blot analysis was carried out using cDNA probes from each of the two EST clusters. When EST 310354 (accession # W30967, representative of the proximal cluster) was used as a probe, transcripts of 1.05, 1.45, 2.4, 6.3, 8.8 and 10.5 kb were found (Figure 3-2B). The use of EST 366663 (accession # AA026167) from the distal region as a Northern probe yielded bands of 1.05, 2.6, 3.6, 8.8 and 10.5 kb (Figure 3-2C). Since a subset of the total number of transcripts was found with each of the two EST cluster probes, this indicates that the two EST clusters represent alternative 3' ends of *IL-17R*.

In order to confirm these results, RT-PCR was carried out using RNA extracted from a CES patient fibroblast cell line (JGe) with various primers from each of the two EST clusters (Figure 3-1B). RT-PCR products were obtained from the following primer sets: IL-F1 & IL-R5 (~1.7 kb), IL-F1 & IL-R6 (~900 bp), IL-F2 & IL-R2 (~800 bp), IL-F2 & IL-R4 (~600 bp), IL-F2 & IL-R5 (~1.6 kb), IL-F2 & IL-R6 (~900 bp), which showed that both ends were represented in transcripts from *IL-17R*. Sequencing of the RT-PCR product obtained from primers IL-F2 and IL-R5 confirmed the link between the 3' end of the coding region and the distal EST cluster. Unfortunately, none of the other bands derived from other primer combinations were sequenced, but since all of the product sizes are too small to represent a read-through, multiple alternative splicing events may be occurring. Since CES patients do not show a phenotype involving the immune system, this gene was not studied further.

Discovery and characterization of CECR1 *variant 2 (*CECR1v2*)*

The *CECR1* gene was discovered within the CES critical region and initially characterized by Dr. Ali Riazi, after which it became the focus of this project. Northern analysis of *CECR1* had revealed two different sized transcripts: 4.4 and 3.5 kb (Riazi et al., 2000), as shown on Figure 1-3. The sequence obtained from the combination of IMAGE clone 54445 and 5'RACE experiments by Dr. Ali Riazi accounted for 3941 bp of sequence (accession # AF190746), suggesting that this transcript corresponded to the 4.4 kb band on the Northern blot (Riazi et al., 2000). The missing sequence (approximately 550 bp) is most probably located in the 5' UTR. To determine the origin of the 3.5 kb band, an analysis of the *CECR1* genomic region was undertaken as described below, which lead to the discovery of an alternative form of *CECR1*.

In order to account for the 3.5 kb band present on the Northern blot, an analysis of ESTs in the NCBI database was undertaken. Within *CECR1* exon 9, a possible alternative polyadenylation (polyA) signal was found at bases 3404-3409. This polyA signal is composed of the sequence ATTAAA, which is found in only about 10% of transcripts (Beaudoing et al., 2000). The sequence just upstream of this possible alternative polyA signal is repeat rich, however there is a unique 97 bp region from bases 2500 to 2596 in the middle of exon 9 of the cDNA. While searching the NCBI EST database with this small unique region of the 3'UTR, a novel EST (IMAGE clone 2190534, accession # AI613429) was found that terminated at base 2793 in the *CECR1* sequence. Interestingly, this EST displayed no canonical polyA signal upstream of its polyA tail, suggesting that either it uses an extremely rare polyA signal, or that this was a spurious but serendipitous transcript. Nonetheless, it was thought that this EST might be an indication of a possible alternate 3' end, and therefore the clone was obtained and sequenced. Analysis of EST 2190534 revealed that the cDNA spanned exons 4 to the middle of exon 9, but had a 70 bp alternative 5' exon, beginning in intron 3 and splicing to exon 4 (Figure 3-3A). This variation of *CECR1* was named "*CECR1* variant 2," or "*CECR1v2*" as opposed to the full length *CECR1* gene, which is therefore referred to as "*CECR1* variant 1," or "*CECR1v1*."

Another sequence described as a "full-insert sequence" (accession # AK074702) was found in the database and subsequently extended the length of the 5' alternative exon to 107 bp. This sequence, in contrast to EST 2190534, contained the full length exon 9 sequence, suggesting that the alternate polyadenylation signal in the middle of exon 9 is not solely used by the *CECR1v2* transcript. In fact, no other variant 2 ESTs have been found that terminate in the middle of exon 9, adding further proof that the EST 2190534 ending was probably not authentic. The total transcript size of *CECR1v2* (based on accession # AK074702) is 3071 bases long, suggesting that it might correspond to the 3.5 kb band previously observed on the Northern blot. This also suggests that there is more sequence (approximately 430 bp) to be found at the 5' end.

In order to test the theory that the 3.5 kb band belonged to the *CECR1v2* transcript, Northern analysis was carried out using two different probes. The 4.4 kb *CECR1* transcript has been shown to be expressed in adult and fetal lung, fetal liver, and placenta, while the 3.5 kb band was expressed in adult heart, pancreas, kidney, and lymphoblast, and fetal lung and kidney (Riazi et al., 2000). Subsequent Northern analysis (within this project) with a PCR probe spanning the entire *CECR1v1* coding region made with primers HID-F1 and HID-R1 (Table 2-1, Figure 3-3A) showed that, in addition to these tissues already shown to express *CECR1*, the 4.4 kb band is also strongly expressed in thymus, spleen, and peripheral blood leukocytes (Figure 3-4A). The smaller 3.5 kb transcript was also faintly expressed in adult and fetal brain, as well as fetal liver. This same blot was probed with a 114 bp ssDNA probe from the 5' alternative starting exon of *CECR1v2*, made by PCR using primers CECR1-Var2-F and CECR1-Var2-R (see Table 2-1 and Figure 3-3A). This probe detected only the 3.5 kb band in adult heart and kidney, and fetal lung and kidney (Figure 3-4B), proving that the smaller (3.5 kb) band on the Northern blot indeed corresponds to *CECR1v2*.

The putative alternative start of the *CECR1v2* transcript creates a predicted protein with 11 unique amino acids followed by the CECR1 sequence coded by exons 4-9, which has been termed "isoform b" (Figure 3-3B) by NCBI staff as indicated in the NCBI record for *CECR1v2* (accession # NM_177405). It is unknown whether the methionine of the predicted ORF represents the actual start codon for variant 2, since there is no upstream stop in the current *CECR1v2* sequence. There is also no predicted signal peptide that would lend support to this methionine being the start codon. The sequence surrounding this methionine is slightly more similar to the Kozak consensus sequence (Kozak, 1996), however, with 6 matches out of 10, compared to 5 for the original *CECR1* gene (*CECR1v1*). A search for other alternative exons upstream of the putative start of *CECR1v2* within introns 1 to 3 of *CECR1v1* using GRAIL2 produced no predicted exons with substantial support (a score greater than 50%). Further characterization of *CECR1v2* will be undertaken by Fang Yang in the McDermid lab, including 5'RACE to determine the full transcript size and genomic structure.

The genomes of close phylogenetic relatives to humans were analyzed for indications of *CECR1v2* using blastn with the sequence of the *CECR1v2* first exon. Both the chimp and baboon genomic sequences contained the alternate starting exon, but all other higher organisms with genomic data, including chicken, zebrafish, and pufferfish (*Takifugu rubripes*) did not. The fact that *CECR1v2* is present only in primates might indicate a specialized function of this variant of *CECR1*. Upon the completion of the pig and cow genomes, it will be possible to determine when variant 2 arose during evolution.

Besides the discovery and Northern analysis of *CECR1v2*, there was no further characterization done with this variant. Therefore for the remainder of the thesis, when a variant is not specified the term *CECR1* implies variant 1. Also, when referring to the genomic sequence, since both variants are implicitly present, the term *CECR1* implies both variants.

Overexpression of human *CECR1* in a transgenic mouse model

In order to study the effect of overexpression of *CECR1* in a model system, transgenic mice expressing human *CECR1* were constructed, through collaboration with Dr. Peter Dickie (HSLAS, University of Alberta). The technique chosen for the creation of *CECR1* transgenic mice was pronuclear injection of the entire human *CECR1* gene contained within a BAC or PAC. The human gene was used to create the transgenic mice, since it would be easier to identify, and because the mouse homologue of *CECR1* had not been found at the time. Later it was discovered that in fact no mouse homologue for human *CECR1* exists (see phylogenetic analysis section).

Characterization of the sequences present on human BAC 609c6

Human PAC 143i13 was not used for the creation of *CECR1* transgenic mice since the putative first exon was localized at the edge of the PAC, and not enough regulatory sequence would be present to allow the natural expression of human *CECR1* in the mouse (Heather McDermid, personal communication). The placement of human BAC 609c6 within the BAC/PAC contig made by Angela Johnson assured that at least part of CECR1 was present (Johnson et al., 1999), although its contents were not fully known since this BAC was not sequenced as part of the chromosome 22 project (Dunham et al., 1999). Southern analysis of BAC 609c6 was therefore carried out within this project to determine if *CECR1* was present in its entirety, as well as to establish if any other genes were present that might compound the phenotype observed in the transgenic mice.

Probes made from various exons and ESTs known to be located proximal or distal to *CECR1* were used in the Southern analysis, based on the known sequence of two adjacent and overlapping PACs (p143i13 and p238m15, see Figure 3-5). EST probes from *CECR6* (EST 46414) and *CECR5* (EST 52444 and EST 1953625) did not hybridize to BAC 609c6, suggesting that these genes are absent. It is possible that some of the 5' end of *CECR5* is located on BAC 609c6, since the 5'-most EST (EST 1953625) does not reach the 5' end of the gene, but a functional CECR5 product could not be expressed in the transgenic mice. The 3'-most exon (Exon 86) of the putative *CECR4* gene does hybridize to BAC 609c6, but this gene has not been completely characterized, and the sequence obtained thus far contains an open reading frame but no start codon (Footz et al., 2001). Since the true 5' end of *CECR4* may or may not be located on BAC 609c6, it is unknown currently whether this gene would be expressed from the BAC. The Exon B probe, which was found through exon trapping by Dr. Ali Riazi (Riazi et al., 2000) and represents the first exon for *CECR1*, did hybridize to BAC 609c6 as expected. Exon38 (within *CECR3*) was also found to hybridize to BAC 609c6, although it is unlikely that a transcript is expressed from *CECR3* since no ESTs have been found for this region, and there is a question as to whether *CECR3* is a pseudogene (Footz et al., 2001).

After the Southern analysis was completed, the end-clones of BAC 609c6 that had been made by Angela Johnson for the construction of the BAC/PAC map (Johnson et al., 1999) were found. Sequencing of the T7 end clone revealed that the centromeric side of BAC 609c6 lies between exons 2 and 3 of *CECR5*, thus confirming that this gene would not be expressed in its entirety. The sequence of the SP6 end clone was unfortunately not of high quality and therefore not useful in determining the location of the telomeric end of BAC 609c6 in relation to the other PACs. However, based on the estimated size of BAC 609c6 (120 kb, Heather McDermid, personal communication), and the length of sequence represented using the above probes, it was estimated that there could only be approximately 10 kb left upstream of *CECR3*. Since nothing else has been predicted within this 10 kb region, and especially since the next gene (*CECR9*) is located 66 kb upstream of *CECR3*, this suggests that no other genes besides *CECR1* are likely to be expressed in their entirety from BAC 609c6. The size of the BAC also ensures that there is ample sequence upstream for the regulatory sequences of *CECR1*.

Production of transgenic founders

The first attempt to create *CECR1* transgenic mice was undertaken by Angela Johnson in the McDermid lab. Out of the 200 injected embryos using human BAC 609c6 at a concentration of 2.5 μg/ml, one founder (Founder 1) was produced from the only litter brought to term. The characterization of the progeny of this founder became part of this thesis. Human BAC 609c6 was purified anew and used for pronuclear injection by Dr. Peter Dickie into mouse embryos. The unusually low success rate in recovering founders from the first round suggested that there might have been a toxic effect of *CECR1* that caused the loss of pregnancies (due to intrauterine death of transgenic pups), therefore lower concentrations were attempted. No births occurred when 1.0 μg/ml of the BAC was used to inject 123 embryos, nor when 0.5 μg/ml was used for 60 embryos. An aborted fetus was found in the cage bedding of one of the four females implanted in the latter round, and when all four females were dissected, various implantation scars and aborted fetuses were discovered in each uterus. It is suspected that if there is a problem with some of the embryos in a pregnancy, the whole litter may be aborted and/or resorbed (Peter Dickie, personal communication). An injection attempt at 0.2 μg/ml brought the total number of unsuccessful injected embryos to 230 for this second round of injections. Finally, embryos were injected at a concentration of 0.1 μg/ml, which resulted in the production of three founders (Founders 2, 3, and 4) from one litter of eight.

Expression of human CECR1 *in the transgenic mice*

In order to confirm the presence of the BAC transgene and to establish that human *CECR1* was being expressed in the lines established from Founders 1-4, Southern and Northern hybridization were carried out. For both analyses, the HID-F1/HID-R1 PCR product was used as a probe. The presence of BAC 609c6 in the initial founders and subsequent mice was routinely determined by Southern analysis of tail biopsy DNA. Northern analysis was carried out by Polly Brinkman-Mills to confirm the expression of the human *CECR1* mRNA in the transgenic line from Founder 1, while RNA from a normal mouse showed no signal (data not shown). Northern analysis of the remaining three lines (Founders 2-4) was completed within this project and is presented in Figure 3-6. Expression of the human *CECR1* transgene was found in all mouse tissues tested. There were two major transcript sizes, measured at approximately 3.8 and 3.0 kb, which were slightly smaller than the expected lengths of 4.4 and 3.5 kb (Figure 3-4). Although faint, the size of the human lung control matched the larger transcript in the mice, suggesting that the difference from the expected sizes was due to differences in Northern blot measurement. The presence of the two transcripts in transgenic mouse brain and heart partly recapitulates the pattern found in humans (Figure 3-4), except that the smaller band does not appear to be present in transgenic mouse kidney. Although the expression of the human *CECR1* transgene was confirmed, determining if the human CECR1 protein was expressed in these transgenic mice was not possible, due to the lack of a good antibody (see below).

Phenotypic observations and mutation analysis of the transgene

The transgenic mice were examined for phenotypes consistent with CES that might be due to the overexpression of human *CECR1*. Each founder produced phenotypically normal, healthy offspring in most cases, displaying no features that could be associated with CES. A few sporadic cases of transgenic pups found dead shortly after birth, and one transgenic runt with a heart defect (enlarged heart due to an anomaly of the aortic valve), were not common enough to implicate the transgene as the cause. The lack of an abnormal phenotype lead to the hypothesis that perhaps a rearrangement, deletion, or point mutation occurred in the BAC before insertion to produce the founders, so an investigation of the DNA and RNA isolated from the transgenic mice was undertaken.

Gross rearrangements of the BAC in each of the four transgenic lines were searched for by Southern analysis. The HID-F1/HID-R1 PCR probe was hybridized to transgenic mouse DNA digested with two different enzymes and compared to the uninjected BAC DNA, but no obvious rearrangements in the transgenic DNA were observed. To uncover small deletions or point mutations, RT-PCR of three segments of the human *CECR1* transcript was carried out using primers HID-F1, -F2, -F3, -R1, -R2, and -R3 (see Table 2-1) on RNA obtained from the four transgenic lines. Attempts were made by the author, Hannah Cheung, and Cheryl Johnson, but no mutations were identified. Therefore, it seems that there were no mutations in the transgene that might prevent the production of the intact CECR1 protein. The presence or absence of the CECR1 protein in these transgenic mice must ultimately be determined using a human CECR1 antibody, to help clarify whether the lack of phenotype is due to the absence of a functional protein. The fact that no mouse *CECR1* homologue exists suggests that overexpression of human *CECR1* might not have an effect on mouse development, although if the CECR1 protein harbours ADA activity like other ADGFs, it might be expected to modulate adenosine levels and therefore have some effect.

Production and analysis of human CECR1 antibodies

Rabbit anti- recombinant human CECR1 antibody production (0A1)

Polyclonal antibodies were raised in rabbits against the bacterially expressed human CECR1 recombinant protein HIDExp1 (see Materials & Methods), in order to characterize human protein expression from the transgenic mice, among other possible experiments. This HIDExp1 construct contains the first third of the putative CECR1 protein, excluding the predicted signal peptide (Figure 3-7). The highly conserved ADA active site residues at the C-terminus were also excluded from this recombinant protein in order to avoid possible cross-reactions with other ADA proteins. Serum from rabbit 0A1 was promising, whereas rabbit 0A6 serum produced no results and was therefore not characterized further.

0A1 serum was used on Western blots of various human and mouse protein tissue lysates. Very faint single or double bands (depending on the tissue) were detected on human samples at sizes consistent with the predicted full length (59 kDa) and mature (without the signal sequence, 56 kDa) CECR1 protein (data not shown). A strong signal was always obtained from the HIDExp1 recombinant protein (which the 0A1 antibody was made against) at approximately 32 kDa, which was close to the expected size (36 kDa) for this positive control. A single dark band (sized approximately 58 kDa) was also detected on a Western blot of the spleen lysates of all three transgenic mouse lines (from Founders 2-4), but this band was also present in the normal mouse lane (Figure 3-8).

This band was still present despite presorbing the 0A1 antibody with bacterial extract, or using antibodies purified on an affinity column. A competition assay was also performed, utilizing increasing amounts of recombinant HIDExp1 protein to compete for the 0A1 antibody and thus decrease the signal from the spleen proteins on a Western blot. Although the signal from the recombinant protein itself was successfully competed away, the human and mouse spleen bands persisted (Figure 3-9), suggesting that they are not specific to CECR1.

Thus although the 0A1 antibody recognized its recombinant protein and seemed to procure the correct sizes for CECR1 in human tissues, the persistent band in the normal mouse lane suggested that the 0A1 antibody might not be specific to the human CECR1 protein. In order to determine the identity of these bands and also to see if the 0A1 antibody was actually able to bind the CECR1 protein, several protein pull-down assays were attempted using various antigen sources and reaction conditions, without success (see below). It was therefore hypothesized that aside from the non-specificity, the 0A1 antibody might not be able to recognize the native CECR1 protein within the pull-down assay, since it was made against a denatured recombinant CECR1 protein. Novel CECR1 antibodies were therefore produced using an alternative method, as described below, in the hopes of being able to detect the CECR1 protein in both native and denaturing conditions.

Rabbit anti- human CECR1 peptide antibody production (2F6, 2F2, and 2F1, 2F3)

In order to obtain an antibody that might recognize both the native and denatured CECR1 protein, anti-peptide antibodies were created. Two peptides were chosen from predicted external regions of CECR1 (Figure 3-7), which were manufactured and injected into four rabbits (2 for each peptide). An ELISA using BSA-conjugated peptide as the antigen showed that each rabbit had mounted an immune response to its respective peptide (data not shown). Western and dot-blot analysis showed that each peptide antibody could detect its associated BSA-peptide conjugate, and that the Pep1 antibodies (2F6 and 2F2) recognized the recombinant HIDExp1 protein as expected. After a long exposure time, a faint ~56 kDa band could be detected in human spleen and patient JGe cell lysates by the 2F6 antibody serum, and very faint ~56 kDa bands were seen in human spleen, heart, and kidney lysates with 2F1 and 2F2 serum (data not shown). These bands, however, were not the only ones observed on the blots, as many other faint bands of various sizes were also seen, suggesting the ~56 kDa bands were not significant.

Protein A pull-down of CECR1 from cell lysates

In order to identify the normal mouse spleen protein detected by the 0A1 antibody (Figure 3-8), and to determine if the 0A1 and/or peptide antibodies were specific to CECR1, several protein A pull-down assays were performed. In theory, the IgG antibodies bound to protein A beads will bind the protein observed on the Western blot. This complex can be precipitated, and the protein eluted and resolved on a gel to be analyzed by mass spectrometry. Various cell sources were tested, including human and mouse spleen, human liver, and human tissue culture cells. Also, various protein extraction methods (denaturing and non-denaturing) and binding times and conditions were attempted, over a period of 1 ½ years, all without success. The 0A1 antibody was used for the majority of these trials, but all four peptide antibodies were also tested. It was thought that perhaps the target protein, CECR1, was not concentrated enough in the ~1mL of cell lysate to be pulled out and detected on a Coomassie stained gel. It is also possible that none of the antibodies used had a high enough avidity to successfully hold CECR1 in order to precipitate the complex.

As a last effort, a large-scale protein A pull-down experiment was attempted, using both the 0A1 and 2F6 serum antibodies and two different sources of protein extract: human spleen and the JGe cell line. The proteins were extracted with the denaturing method for the 0A1 antibody, versus RIPA buffer (non-denaturing) for 2F6. From the 2F6 antibody with the JGe protein lysate, two bands sized approximately 56 and 59 kDa were observed in the extract from the immune-beads on the Coomassie gel. Both bands were excised and sent for Mass Spectrometry analysis at the Institute for Biomolecular Design, University of Alberta. Unfortunately, both samples were identified as vimentin, an intracellular cytoskeletal protein (Clarke and Allan, 2002). There were also IgG peptides identified in the 56 kDa sample. The vimentin protein (accession # A25074) is predicted to have a molecular weight of 54 kDa. The vimentin protein shares one region of strong similarity over the first five residues of CECR1-Pep1 and another region of identity over amino acids 6-8 of the peptide, which together may have caused the cross-reaction. It is unknown why the 59 kDa sample was also identified as vimentin, although perhaps phosphorylation or glycosylation was involved in its size increase.

The faint signals on the Western blots together with the pull-down results with the 2F6 antibody cast doubt on all other antibody experiments, since all antibodies that gave a result seemed to show a similar signal (at 56 and/or 59 kDa). Experimental evidence would be needed to determine if the 0A1 and other peptide antibodies recognize the vimentin protein, however, and instead a new set of CECR1 peptide antibodies is currently being characterized by Fang Yang.

Expression analysis of *CECR1*

In order to gather a developmental profile of *CECR1* expression, and confirm previous expression data, *in situ* hybridization was carried out. Since sections from human embryos are not readily available for expression studies, animal models were evaluated first to narrow down the spatial and temporal expression pattern of the appropriate homologous proteins. Due to the lack of a mouse homologue to human *CECR1*, other model organisms with *CECR1* homologues were relied upon. Preliminary experiments in zebrafish were unsuccessful, so pig was used to gather a developmental profile. The information obtained from the pig was then used as a guide to look at the human expression pattern in a selection of relevant human sections.

Zebrafish whole mount in situ *hybridization*

A 616 bp antisense RNA probe was made from the 3'end of zebrafish *CECR1-1* (there are two zebrafish *CECR1* homologues) in an *in vitro* transcription reaction from the linearized plasmid digested with HpaI. Northern analysis using a ssDNA probe made from this same region showed that *CECR1-1* is expressed in all stages tested (Figure 3-10). Embryos from 14-39 hpf (provided by Dave Pilgrim) were analyzed for the *in situ* experiment, along with embryos at the appropriate stage for the control probe. While the

Krox20 control probe (obtained from Dave Pilgrim) gave the expected staining pattern in the hindbrain (Voiculescu et al., 2001), the zebrafish *CECR1-1* probe did not show any conclusive results beyond some generalized staining in later stages (data not shown). It is possible that the hybridization temperature was too high, or that faint global staining represents the real result. Due to time constraints, these experiments were not continued further.

Pig in situ *hybridization*

Antisense and sense RNA probes were made against three regions of the Pig *CECR1* gene by *in vitro* transcription. The antisense (AS) probe detects the presence of the sense transcript, whereas the sense (S) probe is normally used as a negative control for *in situ* hybridization experiments. The probe made from a restriction fragment of *Eco*RI and *Bsi*WI (probe A, see Figure 3-11A) at the 5' end of the gene gave no signal in preliminary experiments and was therefore not pursued. The remaining probes were made by PCR using primers PID-F1 and PID-R1 (see Table 2-1) for probe 1, and PID-F2 and PID-R2 for probe 2, and the results obtained from these two sets of probes is presented below. Preliminary results with the CECR1-2AS probe (Probe 2, antisense strand that will detect the sense transcript) on day 20 and 28 pig embryo sections gave the same, albeit slightly weaker staining pattern as the CECR1-1AS probe, and therefore this redundant probe was not used in subsequent experiments. As a control, antisense and sense probes were made for the pig myosin heavy chain 2b (*MHC2b*) gene to show staining in muscle tissue. *MHC2b* is expressed in fast glycolytic type 2b muscle fibers, found in both fast and mixed muscles, including skeletal and cardiac muscles (Sterne et al., 1997). PCR primers MHC2b-F1 and MHC2b-R1 (Table 2-1) were designed against the 3'end of the sequence published in GenBank (accession # AB025261) to avoid cross-reaction with the closely related isoforms, *MHC-2a* and *-2x*. RT-PCR using total RNA from adult pig muscle was carried out to yield the insert DNA. Pig embryos ranging from 20 to 31days (Carnegie stages 15 to 22; C15 to C22, approximately equivalent to human day 34 to 55; as observed in the chart on Dr. Mark Hill's webpage, http://embryology.med.unsw.edu.au/OtherEmb/CStages.htm) were used for *in situ* experiments involving both whole-mounts and sections. The best signal to noise ratio was found at a hybridization temperature of 60°C for all of the *CECR1* probes, and 65°C for the *MHC2b* probes.

Pig embryos at 20 days (C15) were used for whole mount *in situ* hybridization. When the MHC2b-AS probe was used, dark staining was observed in the heart atria and ventricles, the myotomes along the spine, and the kidney tubules (Figure 3-12). Although it makes sense that the heart and myotomes were stained with this muscle-specific probe, it is unclear why the kidney tubules were also stained. With the CECR1-1AS probe, generalized staining was observed throughout the embryo, with significant staining in the heart, kidney tubules and spine. The staining in the head might be artifactual, since trapping of the probe in the head often occurs (Rachel Wevrick, personal communication). The staining in the forebrain region of the head does seem significant, however. The negative control, CECR1-1S, showed no staining besides some trapping of the probe in the head.

To further refine the staining patterns observed, sections of day 20 embryos were used for *in situ* hybridization. For the CECR1-1AS probe, generalized staining was observed throughout the day 20 section (Figure 3-13), which corroborates the staining observed in the whole-mount experiment. The kidney tubules, gut epithelia, and patches of the liver (laminae) are stained, along with the ventral part of the head and nose. Faint staining is also observed in the heart (atrium and ventricle) and kidney glomeruli, and the dorsal neural tube may also have some staining. This same staining pattern was observed in the day 28 embryo (Figure 3-14), albeit much fainter. By day 31, the staining in the liver was again patchy, and very faint staining was observed in the heart and kidney (Figure 3-15). The widespread expression of CECR1 was confirmed with RT-PCR using primers PID-F2 & PID-R2, which produced products in all tissues tested, including embryonic (day 28-31) head, heart, liver, kidney, and leg muscle (results not shown).

Use of the CECR1-1S probe showed no staining in any of the three stages, and thus acted as an excellent negative control (Figure 3-13 – 3-15). The CECR1-2S probe was not negative, however, and in fact gave a strong signal in the liver, especially at the day 28 stage (Figure 3-14), which was confirmed by multiple experiments. Staining for the CECR1-2S probe was present in patches of the day 20 liver and in the kidney tubules, but was absent from the heart (Figure 3-13). It is interesting that certain 28 day kidney tubules appeared more darkly stained than others did, since mesonephric tubules are indistinguishable from each other, although they were not as darkly staining as the liver. By day 31, the staining in the liver was quite patchy (Figure 3-15). Although it seemed like the positive patches might be in slightly different locations than the CECR1-1AS probe at this stage, the two sections are not located serially right next to each other. Therefore, it is not known whether the sense and antisense transcripts are expressed in exactly the same liver cells. Since the levels of the transcript detected by the CECR1-2S probe seemed to be declining over time, day 40 sections were checked to determine if its levels were further decreased. There was essentially no staining by the CECR1-1AS probe at this stage, and the staining from the CECR1-2S probe was very faint (data not shown), suggesting that the expression level had in fact diminished further.

The MHC2b-AS probe reliably stained the heart and myotomes in all three stages, but was absent from the gut epithelia (Figures 3-13 – 3-15). The peculiar kidney tubule staining observed in the whole-mounts was again noted in the day 20 section, suggesting that the CECR1-1AS staining in the kidney may be an artifact. On the other hand, there was no staining in the CECR1-1S section, and the incubation temperature used for the MHC2b probe on this section was 60°C instead of the usual 65°C, which may have allowed more background staining. Smooth muscle tissue is present in some parts of the kidney, including the renal pelvis and blood-supplying arterioles associated with the glomeruli (Cormack, 1993), but all the tubules in the section appear to be staining, instead of a small subset. Cranial and dorsal wall muscles were also stained with the MHC2b-AS probe in the day 31 embryo section (Figure 3-15). The MHC2b-S probe was negative for staining in all three embryo stages (data not shown).

Human in situ *hybridization*

To determine the expression profile of human *CECR1*, *in situ* hybridization of human embryonic sections was carried out. Sense and antisense probes were made against four regions of the human *CECR1* transcript, as shown in Figure 3-11B. It was reasoned that if human tissues express a putative antisense transcript encompassing the 3' end of CECR1, region 1 might be expected to have a clean (negative) signal with the sense probe while the region 4 sense probe would not. Probe 1 was isolated by PCR using primers HIDGF-5 and HID-R6, while probe 4 was made using primers HID-F6 and HID-R1 (Table 2-1). It is of note that probe 1 would only detect the variant 1 transcript, while probe 4 would detect both variant 1 and 2. Probes from these two regions were optimized on 10.7-week human fetal liver sections before using the limited set of human embryo slides obtained from the Necker repository in France. The best signal to noise ratio was found at a hybridization temperature of 63°C for *CECR1*-1AS and -1S probes, and 61°C for the *CECR1*-4AS and -4S probes.

In the day 34 (C15) embryo section (approximately equivalent to day 20 in pig) probed with CECR1-1AS, dark staining was observed throughout the liver, and in the excretory tubules of the developing mesonephric kidney (see Figure 3-16, *left*). Very faint staining was also noted in the outer edge of the truncus arteriosis and atrium, and throughout the ventricle of the heart, excluding the endocardial cushions (Figure 3-17, *left*). It is not clear if this result in the heart is real, however, since there was no negative control to compare with for this tissue due to the limited number of slides, and the CECR1-1S embryo section in Figure 3-16 (*right*) is too lateral. Also, since the liver and kidney signals of the CECR1-1AS embryo section in Figure 3-17 (*left*) are much darker than those in Figure 3-16 (*left*), the background staining may be increased in the Figure 3-17 embryo for the CECR1-1AS probe. Dark staining was also associated with the erythrocytes contained within the blood vessels, liver and heart cavities (Figure 3-17, *left*). Staining was absent from the brain, neural tube, eye, and limb buds. The pink hue observed in the spinal ganglia (Figure 3-17, *left*) is an artifact, as it was also observed in the CECR1-1S negative control in Figure 3-16 (*right*).

The CECR1-1S probe was negative for staining in the day 34 embryo (Figure 3-16, *right*). This probe therefore acted as a negative control for the experiment. Unfortunately, the natural colour of the liver is in the same range as the probe signal colour, and therefore it appeared as though there may be staining in the liver when this probe was used. Upon closer inspection, however, the liver did not appear to be stained (Figure 3-16, *middle right*). Structures containing very condensed cells can often give a false positive signal due to a higher background level (Tania Attie-Bitach, Hopital Necker, personal communication). The lumen of some of the kidney tubules appears to have a light pink tinge, which represents background staining in comparison to the

CECR1-1AS probe, in which the entire cell cytoplasm stains a dark pink colour (Figure 3-16, *bottom left*). Therefore, kidney tubule staining was not considered significant unless the pigment was surrounding the nucleus throughout the entire tubule. Staining using the CECR1-1S probe was not evaluated in the heart due to the limited number of sections obtained from the repository.

The CECR1-4AS probe gave a much fainter signal in the day 34 embryo (Figure 3-17, *right*) than that observed for the CECR1-1AS probe (Figure 3-17, *left*), although the pattern of expression was identical, with staining observed in the liver and mesonephric tubules. The staining in the heart was again very faint, and not convincing enough to corroborate the staining observed with the more heavily stained 1AS probe (in Figure 3-17, *left*). In contrast to the CECR1-1S probe, however, the CECR1-4S probe definitely detected a transcript in the liver and kidney tubules of the day 34 embryo (Figure 3-18). The intensity of the stain in the liver and kidney seemed on the same level as the CECR1-1AS (in Figure 3-16) and -4AS probes (in Figure 3-17). The observation of staining with the CECR1-4S probe corroborates the finding with the pig sections, and suggests the possible existence of an antisense transcript in relation to human *CECR1*.

In the day 47 (C19) embryo (approximately equivalent to pig day 25), dark staining was again observed with the CECR1-1AS probe in the liver. Signal was also observed in the adrenal gland, metanephric kidney tubules, pancreas, and gonad (Figure 3-19, *top*). The brain, CNS, heart, lung and stomach were negative. Probe CECR1-1S gave only a very light background staining in the liver and was negative for all other tissues (Figure 3-19, *middle*). The CECR1-4S probe showed moderate liver staining (Figure 3-19, *bottom*), but it was definitely fainter than when this probe was used on day 34 sections, and in comparison to the CECR1-1AS probe on the day 47 section. There was perhaps a very faint stain in the adrenal gland with the CECR1-4S probe. The staining in the remaining tissues for this probe, including the kidney tubules, was negative.

The slides from the 8.5 week (C23) embryo (approximately equivalent to pig day 32) contained a collection of tissues, including heart, neural tissue, lung, metanephric kidney, adrenal gland, and stomach. Use of the CECR1-1AS, or -4AS probe on these slides revealed staining in the fetal (inner) cortex of the adrenal gland (Figure 3-20) and faint staining in the alveolar epithelia of the developing lung (Figure 3-21). The staining of the epithelial lining of the stomach (Figure 3-22) was quite striking with the 1AS and 4AS probes. The staining in these tissues was also observed with the CECR1-4S probe, although the signal was slightly weaker. Staining in all of these tissues (adrenal gland, lung, and stomach) was absent with the 1S probe. There was no staining with any of the probes in the heart or neural tissues at this stage.

Use of the CECR1-1AS or 4AS probe on 10.7-week human fetal liver (no Carnegie stage equivalent) revealed a specific staining pattern (Figure 3-23). Liver cells with large nuclei had cytoplasmic staining, and represent the parenchymal cells of the liver that make up the laminae (Cormack, 1993). Cells with small nuclei were free of probe but counter-stained with methyl green, and are located in the hepatic sinusoids suggesting that they are blood cells. The CECR1-4S probe gave the same pattern, but the staining was much fainter, suggesting the putative antisense transcript was not present in equal amounts to the sense transcript at this stage, while the CECR1-1S probe was negative.

The CECR1-1AS probe stained certain tubules of the 10.7-week metanephric kidney (Figure 3-24A, *left*). The stain was evident in the cytoplasm of cells lining the proximal tubules (Figure 3-24B, *left*), as identified by their brush border and large nuclei (Cormack, 1993). The glomeruli and other tubules were not stained. The yellow and brown staining associated with the glomeruli is likely background staining of the red blood cells, since this was also observed in the CECR1-1S negative control. The kidney sections were otherwise negative when the CECR1-1S probe was used (Figure 3-24A&B, *right*). The CECR1-4AS and CECR1-4S probes gave the same pattern as the CECR1-1AS probe (data not shown), but the staining was much fainter for the CECR1-4S probe.

Confirmation of the antisense transcript

For both the pig and human *in situ* hybridization results, the 5'-most sense probe was negative as expected, but the sense probe at the 3' end of each gene stained very darkly in the liver and kidney tubules. This suggested the possibility of an antisense transcript that might be involved in regulation at the post-transcriptional level. In order to confirm the presence of the antisense transcript, RT-PCR and Northern analysis were performed in both pig and human tissues. Total RNA was extracted from various "adult" pig tissues obtained from a young piglet (through a collaboration with the Pancreatic Islet Transplantation group, University of Alberta), and fetal tissues obtained from day 28-31 embryos (in collaboration with George Foxcroft, University of Alberta). Northern analysis with the *CECR1* ssDNA sense probe 2 (Figure 3-11A) detected antisense bands in adult liver, fetal liver and kidney (Figure 3-25A). The strong band in the fetal liver lane corroborates the strong signal obtained with the pig CECR1-2S probe in the day 28 *in situ* hybridization experiment. The different sized transcripts observed may indicate alternatively spliced products. Use of the ssDNA antisense probe (from the probe A region) to detect the sense transcript showed bands in adult lung and spleen, and perhaps fetal liver and kidney, although the blot quality was not optimal (Figure 3-25B). The presence of the 2.5 kb band corresponds to the 2.2 kb transcript size obtained from sequencing the pig *CECR1* clone, as described in the section on phylogenetic analysis. Although no genomic sequence was available to search for a possible variant 2 transcript in pig, the presence of the smaller 1.7 kb band on the Northern blot suggests that its existence is possible. RT-PCR with primers PID-F2 and PID-R2 (Table 2-1) on pig embryo liver RNA done by undergraduate student Jon Staav also confirmed the presence of an antisense transcript in pig embryos (data not shown). To obtain further sequence of the antisense transcript, and to enhance the proof that the antisense transcript exists, 5' RACE of the Pig *CECR1* antisense transcript was attempted by the author and Jon Staav, but no extra sequence was obtained.

Northern analysis in human also supported the existence of an antisense transcript. A fetal Northern blot (Figure 3-26A) was probed (to detect the antisense transcript) with a 251 bp ssDNA probe that was made against the second half of probe 4 of human *CECR1* (Figure 3-11B) from a PCR product primed with HID-F7 and HID-R1 (see Table 2-1 and Figure 3-27). A single 3.4 kb band was detected in fetal kidney, however no bands were observed in fetal liver. Since the expression of the antisense transcript seems to be modulated over time, the age range of the liver tissue used to make the blot (18-24 weeks, see Figure 3-26) may not be optimal for detection in liver. RT-PCR analysis of human fetal kidney RNA was undertaken, using various primers (HID-F8, HID-I8-A, HID-I8-B, and HID-I8-C, see Table 2-1 and Figure 3-27) within *CECR1* intron 8 as the "reverse" template for the RT reaction. PCR reactions with the HID-R1 "forward" primer produced RT products with both the HID-F8 and HID-I8-A primed RT products, which were sequenced to make sure they were correct, showing that the antisense transcript overlaps with exon 9 and at least some of intron 8 (Figure 3-27). The complete structure of the antisense transcript is unknown, however there were two putative splice donor sites found in intron 8 (Figure 3-27) indicating that the antisense transcript might be spliced elsewhere. A search for ESTs representing the antisense transcript was unsuccessful, and use of the AntiHunter software program to search the *CECR1* genomic region for antisense transcripts yielded no results. 5'RACE will be attempted by Fang Yang, to further characterize the antisense transcript in humans.

Gene structure and expression patterns of the Drosophila *ADGF* genes

The work in this section was published as: **Maier, S.A.**, Podemski, L., Graham, S.W., McDermid, H.E., and Locke, J. (2001) Characterization of the Adenosine Deaminase-related Growth Factor (ADGF) gene family in Drosophila. *Gene* 280: 27-36.

The comparison of a human gene with its homologue in model organisms often aids in gathering more information about that gene. Developmental processes are usually more complex in mammalian systems compared with more simple ones, however similarities in protein sequence are often correlated with similarities in function. Drosophila is an excellent model with which to study the function of a gene, due to the vast genetic tools that have been developed, and the ease of manipulation of these insects. To study the function of *CECR1* therefore, the use of the Drosophila model system was investigated.

Identification and sequencing of genes

Database searches using the tblastn algorithm revealed six *Drosophila melanogaster* genes that had significant amino acid similarity to human *CECR1*, and to each other. The protein products also had significant similarity to human ADA and to *S. peregrina* IDGF, and therefore these genes were labeled Adenosine Deaminase-related Growth Factors (ADGFs) in collaboration with another group working on these genes

(Peter Bryant, personal communication). Five of the six Drosophila homologues (*ADGF-A*, *ADGF-B*, *ADGF-C*, *ADGF-D*, and *ADGF-E*) were discovered as EST clones in the NCBI database. The EST IMAGE clones and accession numbers for each gene were as follows: *ADGF-A* (IMAGE: GH08276; accession # AI109162), *ADGF-B* (AT15281; BF500405), *ADGF-C* (LP05569; AI257258), *ADGF-D* (GH12275; AI134737), and *ADGF-E* (GH18530; AI387855). The clones were obtained through Research Genetics (Invitrogen) and found to be full-length (in that each one has an in-frame, upstream stop) by ABI sequencing using vector and clone-specific primers. Primers GH082-F, GH082-R, and Dros75A-F2 (Table 2-1) were utilized to finish the sequencing of the *ADGF-A* clone, while use of primers GH122-F and GH122-R aided the completion of the *ADGF-D* clone. The full sequence of *ADGF-E* was obtained by cloning restriction fragments into the pGEM-T Easy vector (see Appendix Figure A1) before sequencing with vector primers. These sequences were deposited in the GenBank database with the following accession numbers: *ADGF-A* (AF337554), *ADGF-B* (AF384215), *ADGF-C* (AF337552), *ADGF-D* (AF337553), and *ADGF-E* (AF337551). The sequence of the sixth gene, *ADGF-A2* (AB025255, formerly called MSI) was published by another group (Matsushita et al., 2000).

Genomic structure of the Drosophila homologues

The genomic structure of each Drosophila *ADGF* gene was determined by comparison to the Celera genome sequencing effort (Adams et al., 2000) and/or the Berkeley Drosophila Genome Project (Flybase Consortium, 1999), and is presented in Figure 3-28. The six genes were localized at three different chromosomal locations (51B, 75A, and 87F), according to the genomic sequence descriptions (Figure 3-28A). These locations were confirmed by Lynn Podemski using *in situ* hybridization of BAC probes to salivary gland polytene chromosomes (data not shown). A single gene, *ADGF-E*, maps to 51B. It is composed of two exons and has a predicted transcript length of 1.767 kb (Figure 3-28B). The structure of *ADGF-E* is surprisingly similar to that of *ADGF-B*, which also has two exons. If fact, use of blast-2-sequences to determine amino acid sequence similarity among the six gene products revealed that *ADGF-B* & *-E* are most similar to each other. Two multiple exon genes, *ADGF-C* and *ADGF-D*, localize to 87F and have predicted transcript sizes of 1.78 and 1.63 kb, respectively. They are transcribed in opposite directions with only 1372 bp between their putative transcription start sites, suggesting coordinate regulation of these two genes may occur. Not surprisingly, *ADGF-C* and *ADGF-D* were also shown by blast-2-sequences to have the highest amino acid sequence similarity to each other. Alignment of the two predicted mRNA transcripts using ClustalW and comparison to genomic sequence shows conserved intron locations between these two genes in all cases, suggesting that local chromosomal duplication followed by divergence has occurred. Since the sequence and genomic structure are most similar to each other within *ADGF-B & -E* and *ADGF-C & -D*, the two genes within each pair may perform related functions.

ADGF-A, *ADGF-A2* and *ADGF-B* are all localized within 15.7 kb at chromosomal location 75A. The 1.792 kb *ADGF-A2* transcript is contained entirely within the predicted first intron of *ADGF-A*, and although *ADGF-A & -A2* also appear structurally similar to each, their sequence similarity is not as striking as the other two sets of genes. Genes nested within other genes are quite common in Drosophila, but the nested gene is usually transcribed from the opposite strand (Ashburner et al., 1999). However, since *ADGF-A2* is transcribed in the same direction as *ADGF-A*, this suggested that *ADGF-A2* could be an alternative product from the *ADGF-A* promoter. This was also a possibility for *ADGF-B*. To address these hypotheses, RT-PCR was carried out using primers 75A5'-F1 and 75A5'-F2 (Table 2-1) located in the first exon of *ADGF-A* (the 5' exon), and reverse primers from each other gene in the region: (*ADGF-A*) GH082-R2 and GH082-R3; (*ADGF-A2*) MSI-R3 and MSI-R4; (*ADGF-B*) 75L-R2 and 75L-R3 (Figure 3-29A). RT-PCR products were found joining the 5'exon with the *ADGF-A* gene, as expected based on the cDNA clone sequence (GH08276), but no products were obtained for *ADGF-B*, indicating that the first exon of *ADGF-A* is not shared with *ADGF-B*. For the *ADGF-A2* gene, splice products were indeed obtained from the 5'exon (Figure 3-29B), although the bands were very faint and required amplification. Sequencing of these products confirmed the they were the result of alternative splice products from the donor splice site of the 5' exon of *ADGF-A* to either 12 or 18 bp downstream from the published start codon in exon 1 of *ADGF-A2* (Matsushita et al., 2000). Thus, two alternative acceptor sites located downstream of the published start codon appear to be used, although both new transcripts maintain the exact same ORF when translated. The next start codon is located downstream of both splice acceptor sites, which excludes 19 amino acids from the N-terminus of the published ADGF-A2 protein and therefore produces the same protein for either splice form. Interestingly, neither splice variant contains an in-frame upstream stop, while the published transcript does. The fact that the 5'exon is shared between *ADGF-A & -A2* suggests that coordinate regulation also exists between these two genes, as was suggested for the *ADGF-C & -D* set.

Conservation of intron positions

Due to the structural similarity among these six genes, it was suspected that the intron positions were conserved among some of the structures. An alignment of the six ADGF predicted protein products, along with human CECR1, was performed using ClustalW and the intron locations were placed on the alignment, as compared to genomic sequence. Figure 3-30 illustrates the gene structures along with the intron locations shared between two or more genes. Among the Drosophila genes, all of the intron locations are shared among at least two genes, which further adds to the structural similarity within the three gene sets: *ADGF-B* to *ADGF-E*, *ADGF-C* to *ADGF-D*, and *ADGF-A* to *ADGF-A2*. One intron position is shifted by one base between the *ADGF-E*, *-B*, *-A*, & *-A2* genes, and the *ADGF-C & -D* genes, which may have arisen through the process of intron-sliding (Rogozin et al., 2000). This similarity of the intron/exon structure among the Drosophila genes indicates a common origin of all six. Surprisingly, two Drosophila intron positions are also shared with the human *CECR1* gene, suggesting that these intron positions may have been present in a common ancestor. This issue was addressed more completely using the full phylogenetic analysis described later in this chapter.

Expression analysis of the six Drosophila ADGF *genes*

The presence of six Drosophila *ADGF* paralogues compared with only one human gene, combined with the structural and sequence similarities comprising the three gene sets, suggested the possibility of redundant function between the Drosophila genes. In order to test this hypothesis, the developmental expression patterns of each Drosophila *ADGF* gene was determined using Northern analysis and RT-PCR. Northern blots were prepared from poly(A)+ mRNA isolated from Drosophila embryos, larvae, pupae, adult males and adult females by Lynn Podemski. Successive hybridizations were carried out by Lynn Podemski using DNA probes (described in the legend of Figure 3-31) for each of the six *ADGF* genes. A probe against the *rp49* gene was used as a control for all the blots. RT-PCR was carried out by the author using total RNA from the same five Drosophila tissue samples to confirm the Northern blot analysis and detect any low level expression (Figure 3-31, beneath the Northern blot for each gene). The primers for each gene were designed across introns and are listed in the legend of Figure 3-31, and in Table 2-1. *RPS3a* served as a control for all stages.

For each gene, the mRNA transcript size corresponded to the cDNA size, suggesting that the cDNAs are approximately full length and that the predicted gene structure is correct. Both *ADGF-A* and *ADGF-E* detected two bands on the Northern blot (Figure 3-31). In both cases the larger band correlates to the size of the cDNA clone, while the smaller band may represent alternate splicing or mRNA processing, but this was not confirmed. The six Drosophila genes showed different developmental expression patterns (Figure 3-31). *ADGF-B* and *ADGF-A2* appear to be male specific by both Northern analysis and RT-PCR, although male-specificity in larva and pupa is only inferred. *ADGF-C* and *ADGF-E* seem to be male predominant, since by Northern analysis they appear male-specific, but are present in both adult sexes by RT-PCR. *ADGF-A* and *ADGF-D* are more universally expressed, being detected by RT-PCR in all stages and both sexes. None of the other four genes (*ADGF-A2*, *-B*, *-C*, and *-E*) show expression in embryo, and each has low expression in larva (i.e. only detected by RT-PCR). These expression patterns have been confirmed independently (Matsushita et al., 2000; Zurovec et al., 2002). RT-PCR was also carried out for Drosophila "ADA", using primers 85C-F and 85C-R (Table 2-1), and bands were observed in adult, pupa, which were the only stages tested.

Thus, it seems that the three sets of genes that were grouped together by sequence and structural similarity have been separated into single functional genes by differential expression patterns. Although since there is still a great deal of overlap in the expression patterns of the six Drosophila *ADGF* genes, some redundancy is still likely. In order to study the possible redundancy between genes, mutations in one or more of the *ADGF* genes would be necessary. A search of mutant fly stocks (P-element insertions or deletions) was unsuccessful, therefore experiments analyzing the loss of one or more genes in combination were not carried out. The possible redundancy in function of these genes makes it more difficult than previously thought to study the function of human *CECR1* homologues in the Drosophila model system, and therefore this work was not continued further.

Phylogenetic analysis of the ADGF subfamily

The data in this section was published as: **Maier, S.A.**, Galellis, J.R., and McDermid, H.E. (2005) Phylogenetic analysis reveals a novel protein family closely related to adenosine deaminase. *Journal of Molecular Evolution* (in press).

From the discovery and analysis of the six Drosophila *ADGF* genes, and in comparison to human *CECR1*, it was clear that together these genes were very complex, and might be part of a larger family of interesting and perhaps developmentally important genes. As mentioned in the introduction, many genes were published throughout the duration of this project with similarity to human CECR1 and ADA. Note that the term "CECR1" is used for vertebrate homologues that hold membership within the larger ADGF subfamily. The compilation of all possible genes with sequence similarity to ADA and ADGF would allow evolutionary and perhaps functional relationships to be forged. Therefore, a phylogenetic analysis was undertaken with as many homologous sequences that could be gleaned from the database at the time.

Identification and sequencing of preliminary protein sequences

Some of the *ADGF* homologues were uncovered by the author of this thesis in conjunction with other members of the McDermid lab, especially the *CECR1* homologues in vertebrates. A pig (*Sus scrofa*) *CECR1* cDNA clone (accession # F14844, kindly provided by A.K. Winteroe, Denmark) was sequenced with the help of Polly Brinkman-Mills, and published (accession # AF384216).

A zebrafish (*Danio rerio*) EST (accession # AW077621) with similarity to *CECR1* was discovered by a tblastn search, but it was unavailable for ordering. Primers ZID-F1 and ZID-R1 (Table 2-1) were designed based on the EST sequence, and a PCR probe was made by RT-PCR from mixed-stage zebrafish RNA obtained from Angela Manning, University of Alberta. The PCR probe was used to screen a 19-25 hpf (hours post fertilization) cDNA library, and the one positive plaque obtained was excised and sequenced using primers ZID-F1, ZID-F2, ZID-F3, and ZID-R1 along with vector primers to obtain the full-length zebrafish *CECR1-1* clone (2.5kb). The sequence of this clone was submitted to the database (accession # AF384217).

A *Xenopus laevis* EST clone (accession # BJ040131) with amino acid similarity to human CECR1 was found in the NCBI database and obtained from N. Ueno, National Institute for Basic Biology, Okazaki, Japan. Sequencing with XenoCECR1-F1, XenoCECR1-R1 (Table 2-1), and vector primers revealed that the full-length gene was present except for a gap in the sequence corresponding to human *CECR1* exon 3. Julia Galellis used RT-PCR to show that this gap was likely a cloning artifact, and replaced the gap in the sequence, using primers XenoCECR1-F3 and XenoCECR1-R3 (Table 2-1) on adult *Xenopus laevis* spleen RNA. This full length sequence was submitted to the database (accession # AY902778).

A full-length chicken (*Gallus gallus*) EST clone (accession # CD738959), with similarity to human CECR1 was obtained from H. Lillehoj, Animal Parasitic Diseases Laboratory, Beltsville, Maryland, USA, sequenced by Fang Yang, and submitted to the database (accession # AY902779).

No mouse CECR1 homologue exists

Various approaches were attempted in order to find the mouse homologue to human *CECR1*. Low stringency Southern (Dana Shkolny) and Northern (Polly Brinkman-Mills) analysis showed faint but promising bands when human *CECR1* exon 1 (Exon B) was used as a probe. However, mouse cDNA (Polly Brinkman-Mills) and BAC library screens (Jennifer Skaug, TCAG, Toronto), degenerate PCR, and searches of the mouse cDNA and EST databases gave no results. Blast searches of the nearly complete mouse and rat genomes were also unsuccessful. The lack of a rodent homologue in a gene family has been observed elsewhere (Lutz et al., 1994; Fougerousse et al., 2000) and so it is possible that no rodent homologue exists for *CECR1*. Mouse BAC 541L22 was thought to contain mouse *Cecr1*, since Southern analysis using human probes allowed Tim Footz to confirm that the mouse homologues of the human genes surrounding *CECR1* were present on this BAC (Footz et al., 2001). When the complete sequence for BAC 541L22 became available, however, it was evident that only a remnant of exon 1 and part of intron 3 were present, suggesting that the rest of mouse *Cecr1* was deleted or lost. All of this evidence together suggests that no mouse equivalent to human *CECR1* exists.

Discovery of the ADAL paralogues

While searching for a mouse homologue for human *CECR1*, a full-length mouse EST (accession # BC052048) was discovered with slight protein similarity (40%) to the C-terminal region of the CECR1 protein. This mouse protein showed slightly more similarity (41% over the entire length of the protein) to ADA, and was therefore termed *Mus musculus* Adenosine Deaminase-Like (ADAL). The clone was obtained from Open Biosystems and the sequence, as deposited in the database, was confirmed by Rezika Zurch and Twila Yobb. A human *ADAL* homologue was discovered on chromosome 15 and its expression was confirmed by RT-PCR (Melanie Kardel and Nic Fairbridge, unpublished results). The discovery of these two ADAL proteins spearheaded the discovery of ADAL homologues in various other organisms through database searches by the author. The databases were also scoured for novel ADGF homologues. ADA protein sequences were collected *in silico* from organisms with ADGF or ADAL representatives, in order to make the phylogenetic analysis more complete. Therefore, there are three distinct protein subfamilies with significant sequence similarity to each other: ADGF, ADAL, and ADA.

Identification of protein sequences in silico *for use in the phylogenetic analysis*

The human proteins of these three subfamilies were used separately to find putative orthologous genes and their corresponding protein products from all available organisms from the GenBank databases (see Methods). As they were discovered, gene products were named according to sequence similarity to other proteins and/or numbered in the order in which they were found, and were included in Table 3-1. In most cases, the protein sequence of each gene was predicted from the first methionine within the open reading frame after an upstream stop codon, or by comparison to the predicted start of the closest homologue, where no upstream stop was found. This collection of proteins was finalized in August 2004.

Some cDNA sequences were found in their entirety in the database and did not require any perturbation. The human ADA gene (accession # NM_000022) was found by searching the PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) nucleotide database with the words "human ADA." The mouse, rat, Xenopus, zebrafish, and *C. elegans* ADA genes were discovered by a tblastn search of human ADA against the nr database. One ADAL representative in *Xenopus laevis* was also discovered as a complete clone, and was named Xl_ADAL (accession # BC073685).

Many genes were predicted *de novo* from genomic DNA by comparison to the respective human protein, gene prediction, and/or manual extraction of the nucleotide sequence and assembly based on the tblastn result (labeled A in Table 3-1, under the Gene accession and Protein accession headings). *P. troglodytes* CECR1, ADAL, and ADA, *P. anubis* CECR1, *G. gallus* ADA and ADAL, *D. rerio* CECR1-2, *T. rubripes* CECR1-1 & -2, ADAL, and ADA, and *T. nigroviridis* CECR1-1 & -2 were all predicted in this manner. A gap in the genomic sequence of *T. rubripes CECR1-2* was closed by PCR and sequencing using primers Fugu2-F1, Fugu2-F2, and Fugu2-R1 (see Table 2-1) with DNA obtained from the UK HGMP Resource Centre, Cambridge. *D. pseudoobscura* ADGF-A, -A2, -B, -C, -D, & -E, and *D. yakuba* ADGF-A, -A2, -B, -C, & -D also fall into this category of predicted genes, and are so named due to their close similarity to their respective *D. melanogaster* genes. The *D. pseudoobscura* and *D. yakuba* ADAL genes were also predicted from genomic sequence, by comparison to the *D. melanogaster* "ADA" gene. Note that "ADA" is the official FlyBase name for this protein in *D. melanogaster*, but it is more closely related to the ADAL family by blast searches. A fragment of an ADGF homologue was found in the genomic sequence of *Oryza sativa* (rice) but not enough sequence was present to predict a complete gene, so this gene was not pursued further.

Some putative proteins were already predicted from genomic data by genome curators, and placed in the database. A subset of these appeared to be correctly predicted, including *D. melanogaster* "ADA," *A. gambiae* ADGF-1, *C. elegans* ADAL, *U. maydis* ADAL, *N. crassa* ADGF-1 & -2, *G. zeae* ADGF-1, *M. grisea* ADGF-1, *A. nidulans* ADGF-1, and *D. discoideum* ADGF (labeled C in Table 3-1, under the Gene accession and Protein accession headings). Another subset of predicted proteins appeared not to be entirely correct, based on comparison to other subfamily members, and were altered using an assembly of EST data, GENSCAN predictions, and/or by subfamily comparison in order to obtain a better prediction. Proteins in this category were labeled B in Table 3-1 (under the Gene accession and Protein accession headings) and include *R. norvegicus* ADAL, *T. nigroviridis* ADAL and ADA, *A. gambiae* ADGF-2 & -3 and ADAL, *U. maydis* ADGF, *N. crassa* ADAL, *G. zeae* ADGF-2, *M. grisea* ADGF-2, and *A. nidulans* ADGF-2 and ADAL.

Where possible, EST sequences were obtained to lend proof of expression of the predicted genes, using tblastn of the predicted protein against a species-specific EST database. Some of the putative genes predicted from genomic DNA had at least one EST in the database to support the existence of the gene, including *G. gallus* CECR1, ADAL, and ADA, *T. rubripes* ADA, *D. melanogaster* ADA, *A. gambiae* ADGF-1 & -3, *C. elegans* ADAL, *N. crassa* ADGF-1 & -2, *G. zeae* ADGF-1, *A. nidulans* ADAL, and *D. discoideum* ADGF. Although there were no ESTs in the database, the expression of the *D. rerio* CECR1-2 gene was confirmed through RT-PCR by Fang Yang, although the full sequence has not yet been obtained. The rest of the predicted genes, however, remained simply predictions that had no ESTs in the database to corroborate the gene's existence (see Table 3-1). In some cases, this might have been due to a lack of any ESTs available for a certain organism. For example no ESTs for *T. rubripes* had been entered into the database as of August 2004, therefore it could not be determined whether or not the two predicted *T. rubripes* CECR1 homologues were expressed. In other cases, some ESTs had been found for other genes from the same organism, and therefore the lack of an EST for a certain gene could be more informative. For example, at least one EST was present in the database for both the *A. gambiae* ADGF-1 & -3 genes, but none were found to corroborate the existence of the *A. gambiae* ADGF-2 gene. If many ESTs representing all tissues and time points had been entered for *A. gambiae*, this might indicate that the ADGF-2 gene is not expressed, however if only a few ESTs had been entered, perhaps an EST to represent this gene had just not been found and deposited in the database yet.

Only sequences that could be predicted in their entirety were included in the analysis. Significant similarity to one of the three families was observed in the following protein homologues, however they could not be predicted in their entirety and were therefore marked with a # symbol within Table 3-1 and omitted from the phylogenetic analysis: *S. scrofa* ADAL and ADA, *D. rerio* ADAL, *T. rubripes* ADA, *D. yakuba* ADGF-E, *A. gambiae* ADGF-4, and *G. zeae* ADAL.

Prediction of signal peptides

Some members of the ADGF subfamily have been shown to be secreted (see Introduction). Therefore, the predicted cellular location of all members of the ADGF, ADAL and ADA protein subfamilies was determined, using signal and/or cellular localization prediction software. Most of the ADGF members were predicted to have a signal peptide, and the likely residues cut off during secretion are indicated in Table 3-1 (under the column labeled SP). Among the Drosophila species, the ADGF-A, -C, & -D proteins were all predicted to have a signal sequence, whereas ADGF-A2, -B, & -E were not. The Drosophila ADGF-B & -E proteins instead were predicted to be targeted to the mitochondria, a fact that is further strengthened by the genomic structural similarities shared between the Drosophila *ADGF-B & -E* genes. Drosophila ADGF-A2 is suspected to be a transmembrane protein (Matsushita et al., 2000), while the predicted cytological location of the *A. gambiae* ADGF-1 & -3 could not be determined. The *D. discoideum* ADGF protein was predicted to contain a signal peptide whereas the fungal ADGF proteins were not, perhaps because the fungal organisms exist as single cells. In *D. discoideum*, the amoeboid cells aggregate and can form a multi-cellular fruiting body during starvation conditions (Weijer, 2004).

As expected, none of the ADA proteins were predicted to contain a signal sequence, since ADA is a cytosolic protein (Franco et al., 1998). Also, none of the ADAL proteins were predicted to contain a signal sequence, suggesting perhaps that this group of proteins may be more closely related to the ADA subfamily than the ADGFs.

CANADIAN THESIS
THESES CANADIENNES

MAIER, STEPHANIE ANN
USE OF MODEL ORGANISMS AND
PHYLOGENETIC ANALYSIS TO ...

ISBN 0-494-08688-2
NR-08688
2 of 3

Alignment of protein sequences

In order to address whether the ADGF, ADAL and ADA gene subfamilies were evolutionarily related, several phylogenetic analyses were undertaken. Since the DNA sequences showed no significant similarity between the three subfamilies, the putative protein products were compared. Also, because adenine deaminase (ADE) and AMP deaminase (AMPD) share a common reaction mechanism with ADA (Becerra and Lazcano, 1998), several representative members of these two subfamilies were included in the phylogenetic analysis, to better resolve the inferred tree (see Table 3-1). Since two groups of adenine deaminases have evolved independently from two different ancestral proteins (Ribard et al., 2003), only the group of ADEs with sequence homology to the ADA subfamily was used for the phylogenetic analysis. *E. coli* ADE belongs to the group that does not share sequence similarity with ADA, and therefore does not appear in the analysis. Although there were several vertebrate AMPDs discovered in the database, only prokaryotes and fungi possess ADE (Ribard et al., 2003). The definition of subfamily and family has been outlined previously (Riveros-Rosas et al., 2003) and therefore the ADGF, ADAL, ADA, ADE and AMPD subfamilies are described as belonging to the adenyl-deaminase family.

An initial MUSCLE alignment was constructed with all 95 protein sequences from the five subfamilies listed in Table 3-1. There were eight highly conserved regions found throughout the alignment of the five subfamilies, mainly focused around the catalytic residues required for ADA activity. A region was included if it was composed of at least three contiguous conserved residues, with at least two of the residues showing conservation in most members of at least three subfamilies. In order to focus on these eight important regions, the conserved amino acid residues were shaded by BOXSHADE and presented in Figure 3-32. Since functional importance is highly correlated with evolutionary conservation (Gu, 2001), the residues that are conserved amongst the five different subfamilies might indicate functional importance for the deamination process. The phylogenetic analysis (presented below) showed that the AMPD subfamily was a natural outgroup of the four remaining groups, and the following observations are discussed in light of this fact. It was stated in the introduction that the crystal structure of mouse ADA revealed the residues important for ADA function: His15, His17, Gly184, His214, Glu217, His238, Asp295, and Asp296 (Wilson et al., 1991). For simplicity, all residue numbers discussed hereafter within the alignment refer to amino acid positions within the mouse ADA protein sequence, unless otherwise stated.

Within the first domain, the ADGF subfamily shares a motif consisting of methionine (or iso/leucine), proline, lysine and glycine (MPKG), the beginning of which corresponds to position 9 in the mouse ADA protein. The ADAL and ADE proteins share a conserved leucine or methionine in the first position, and both the PK residues, but not the glycine in the fourth position. The ADA proteins only conserve the PK residues of this motif, except *E. coli* ADA. The conservation of the proline and lysine residues throughout the ingroup suggests that these residues are important for the function of these proteins, but their role in ADA activity has not been demonstrated (Wilson et al., 1991; Sideraki et al., 1996; Mohamedali et al., 1996). The fact that the glycine is common only to the ADGF proteins suggests that it may perform a critical function only in this subfamily. The AMPD subfamily seems to have retained some remnants of the full MPKG motif, but this domain was clearly not conserved over time in this group. The two ADA active site residues, His15 and His17, located at the end of conserved domain 1 are almost completely conserved among all ingroup proteins, but not within the AMPD outgroup. These two histidines are thought to be important for zinc binding (Wilson et al., 1991; Mohamedali et al., 1996). The leucine residue just previous to these important histidines is also mostly conserved throughout the ingroup, with the exception of the fungal ADEs, suggesting it may be important as well. Again, only remnants of these three residues are observed in the outgroup, suggesting they were not important in the function of this subfamily. Asp19 is not conserved in proteins outside the ADA subfamily, although it was suggested to be important in the activity of ADA (Wilson et al., 1991), and was therefore not included within domain 1.

Domain 2 within the ingroup consists of a total of nine residues; five conserved residues alternating with four less conserved sites (Figure 3-32). The last two alternating residues, glutamate (Glu; E) and arginine (Arg; R), are conserved within the entire alignment, except where Arg was changed to phenylalanine (Phe; F) for the ADEs. This last Arg residue of this domain corresponds to Arg101 in the mouse ADA protein, which is thought to form a salt bridge with Glu260, an interaction that may be important for stability (Wilson et al., 1991). The third domain is generally conserved within the ADGF, ADAL, and ADA groups. The Gly184 residue that is important for ADA activity (Wilson et al., 1991; Sideraki et al., 1996; Mohamedali et al., 1996) is completely conserved throughout all three subgroups, except the three Drosophila ADGF-A2 proteins. Instead of the glycine residue in this position, the ADE proteins have a serine (Ser; S) and while some AMPDs have a serine in this position, others do not. Since this residue is not conserved in the AMPD and ADE groups, it might be important only for the adenosine substrate, although this has not been confirmed. The fourth and fifth domains are generally conserved throughout the entire alignment, although conservation in the AMPDs (for both domains) and the ADALs (for domain 4) is less strict. The three important residues within these two domains, His214, Glu217 and His238, are highly conserved with only a few exceptions in some of the insect ADGFs. But since the insects seem to have an over-abundance of ADGF proteins (the Drosophila species harbour six ADGF proteins, and *A. gambiae* has at least four ADGFs), this suggests that perhaps not all of these proteins are functional, or that some paralogues might have a different activity. Indeed *D. melanogaster* ADGF-E has previously been described as lacking ADA activity (Zurovec et al., 2002). Domain 6 is composed of a number of residues that are highly conserved throughout the entire alignment. Particularly, Glu260 and Ser265 have been suggested to form salt bridges with Arg101 and His238, respectively (Wilson et al., 1991). Glu260 is conserved in all ADGF, ADAL and ADA proteins. Ser265 is conserved in every sequence except in two fungal ADALs, but these two proteins share a serine residue one position upstream, which may perform the same function. The seventh domain consists only of two important ADA active site residues, Asp295 and Asp296, and a proline (P) generally conserved throughout the alignment except for the ADALs. The final domain begins at mouse ADA residue 325, and although it is conserved more within the AMPD and ADGF subfamilies, its relevance is not known.

Initial phylogenetic inference

An initial Bayesian analysis was performed using the alignment containing all five protein subfamilies, and the consensus tree was large and complex, due to the number of taxa involved. A simplified version of the tree was constructed by removing individual taxa from the tree to leave the overall relationship between the five protein subgroups. As shown in Figure 3-33, the ADAL proteins clearly form a cluster with the ADA and ADE subgroups, although much phylogenetic change has occurred between the latter groups, as represented by the long branch connecting the ADEs to their common ancestor. The phylogenetic relationship of ADA and ADE has already been established in the literature (Ribard et al., 2003), but the ADAL subfamily is a novel addition. The ADGF subfamily is distantly related to the previously mentioned groups, but the AMPD members are most distant, noted by the very long branch connecting them to the other groups. This indicated that the AMPD subfamily seemed to be a natural outgroup to all the other proteins and allowed the tree to be rooted from the node that the AMPDs originated from (see below). Note also that the tree topology correlates with the size of the proteins. The AMPDs have an average amino acid length of 746 (+/- 64, standard deviation), the ADGF subfamily had an average length of 531 (+/- 34), while the ADAL, ADE, and ADA groups had lengths of 351 (+/- 9), 351 (+/- 10), and 359 (+/- 15), respectively. Based on the assumption that the AMPD subfamily was the outgroup, and due to the added complexity in the alignment when these larger proteins were included, the AMPD subfamily was excluded from further in-depth analyses of the ingroup.

Focused analysis of the ingroup and aspects of MrBayes analyses

A second MUSCLE alignment of the 80 protein sequences belonging to the ingroup (ADGF, ADAL, ADA, and ADE) was used for further phylogenetic analyses. A series of initial (150,000 generations, trials 1-9) MrBayes runs were conducted and the same basic topology was recovered each time, except for slight changes in the clades with low support values. The difference between these runs was the temperature setting that dictates the amount of change to the posterior probability, and the resulting acceptance rates for swaps between chains. The acceptance rates summarize the number of times a swap occurred between chains separated by only one heating step. According to the MrBayes tutorial (http://workshop.molecularevolution.org/software/mrbayes/) (Ronquist and Huelsenbeck, 2003), the acceptance rates should lie between 10-70%. The first trial, using the default temperature of 0.2, produced acceptance values in the appropriate range. A second trial was performed to confirm the results of the first trial. Although the topology was identical, one of the three temperature acceptance values was not between 10-70% (see Table 3-2), indicating that one chain was not used for swapping. The MrBayes tutorial manual suggested that if the acceptance rates were too low, the temperature setting should be lowered in an attempt to increase the acceptance of chain swaps.

As shown in Table 3-2, many different temperature settings were evaluated. None of the temperatures evaluated in the initial analyses, besides the first one, produced a set of acceptance values with all numbers lying in the correct range. But, as stated above, the resulting tree topology of all of these trials was basically identical, except for the extreme temperature of 0.001 in trial 6 that produced a major polytomy (set of collapsed branches) within the amphibian and fish ADAs. Also, besides the first trial, half of the subsequent trials with a temperature of 0.2 had two of the three values in the acceptable range, suggesting that of all the temperatures evaluated, perhaps a temperature of 0.2 was the best value to use for this data set. All taken together, in trials where not all of the four chains within one run were used to produce samples for the tree, the same topology was generated compared to trials that did use all four chains to produce the final consensus tree. Therefore, changes in temperature and the resulting acceptance of swaps between chains did not affect the outcome of the tree topology for this data set.

The final Bayesian analysis (550,000 generations, trial 10) was run 5 times (runs A-E), using a temperature setting of 0.2, and the resulting tree topology was identical each time. Although two of the five runs had swap acceptance values outside of the appropriate range (see Table 3-2), all five analyses were included in the results since at least two chains per run were sampled from in order to give the final output. Also, the increase in the number of generations seemed to help with the acceptance values, since three of the five final runs had all three acceptance values within the range 10-70%. For each of the five runs (A-E), the convergence to a stationary likelihood value happened very quickly (before 30,000 generations) although each run seemed to take a slightly different path (Figure 3-34). To be safe though, the first 50,000 generations were discarded as burn-in. The first analysis (A) was chosen to be a representative of the five runs, and its associated branch lengths and posterior probabilities are presented in Figure 3-35.

Between the five MrBayes runs, the posterior probabilities (support values) at each node varied between the five runs by up to 5% in most cases, which was expected with this sampling methodology. Support values for three nodes fell outside the 5% standard deviation range (data not shown). The node depicting the common ancestor of the insect, vertebrate and *U. maydis* ADALs had a standard deviation of 11% while the clade internal to that, containing the vertebrate ADALs and *U. maydis* ADAL, had a standard deviation of 7%. Support for the first node was 83%, while for the second it was only 63% on the representative tree (Figure 3-35). This indicates that perhaps each of the five runs had a disagreement in the number of trees sampled with *U. maydis* ADAL basal to the vertebrate ADALs, and questions whether this fungal gene should even be placed within the ADAL deuterostomes. The third node with a standard deviation greater than 5% was that connecting *X. laevis* ADA to the avian and mammalian ADA proteins. The support for this node between the five runs varied from 85% to 100%, and resulted in a standard deviation of 6%. Perhaps these values would have stabilized to a better consensus if the chains had been run longer.

Observations from the Bayesian analysis

Two major clades were evident in this tree: 1) ADAL, ADE and ADA, and 2) ADGF (Figure 3-35). This major split between the two groups was well supported, as indicated by the posterior probability value of 1.00 (represented as a percentage out of 100% in the figure). This major split was also observed in the initial analysis that included the AMPD protein sequences, and the approximate placement of the "ROOT" between these two groups in Figure 3-35 represents this outgroup. Overall, the entire topology was well supported, especially for many of the deep divergences, with only a few weak posterior probabilities for some of the internal nodes.

Within the first major clade, support was high for the split between the ADAL and ADE/ADA groups (100 and 99%, respectively), but support for some internal nodes within each group was problematic. Bayesian analysis tends to give high posterior probabilities for internal nodes, such that a value less than 70% might be considered to be low (Huelsenbeck et al., 2001). Support values for the nodes leading to *C. elegans* ADAL (shown as a polytomy, or collapsed branch, in Figure 3-35), *U. maydis* ADAL, and the two pufferfish ADALs (*T. rubripes* and *T. nigroviridis*) were <50%, 63%, and 64% respectively. This suggested that although these proteins clearly belong to the ADAL subfamily, there was a lack of confidence for their placement within the subfamily. Also, the placement of these taxa disagrees with the accepted phylogenetic relationship of organisms (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=taxonomy) known to date, which shows that *U. maydis* ADAL should form a clade with the two other fungal ADALs, and the pufferfish ADALs should appear basal to *X. laevis* ADAL. Support for the inclusion of *E. coli* ADA in a monophyletic group with the other ADAs was only 51%, as this protein sequence was often grouped with the ADE subfamily. Support was also less than optimal in clades containing *T. nigroviridis* ADA and *G. gallus* ADA, since the posterior probabilities for these clades was in the low 70's.

Within the second major clade, the ADGF proteins fall into several distinct subgroups that were very well supported: fungi, vertebrates, and insects. *A. californica* MDGF appears basal to the insects, while *D. discoideum* ADGF is basal to both the vertebrates and insects. The general scheme of this major clade agrees with the accepted organismal phylogeny, except that *D. discoideum* should appear basal to the entire group, yet this bipartition was only found in 6% of the sampled trees. Note that there are several duplication events within this major clade that occurred after these groups branched from the common phylogenetic tree. All the fungi seem to have two ADGF homologues, except *U. maydis*, which either has lost one copy or it has yet to be sampled. Also, there is a discrepancy as to whether *A. nidulans* ADGF-1 should appear basal to the other fungal ADGF-1 genes as shown, with a low associated support value of 71%, or whether it should form a monophyletic clade with *A. nidulans* ADGF-2 (15%), or even switch places with it (14%). The fish *CECR1* genes have also undergone a further subdivision, which may be a result of the tetraploidation of ray-finned fish (reviewed in Taylor et al., 2003). Instead of forming a clade with the other fish CECR1-1 proteins, *D. rerio* CECR1-1 appeared basal to the entire group, indicating that perhaps this gene has retained more of the ancestral features than the other genes.

Within the insect subgroup, the six Drosophila ADGF gene products act as a backbone onto which the other insects with less fully sequenced genomes may be placed. For example, only one ADGF family member has been discovered in the flesh fly, *Sarcophaga peregrina*, and this protein groups with the Drosophila ADGF-A members. Sequencing of the entire *S. peregrina* genome may reveal five other gene products similar to the other ADGF members present in the Drosophila species. There was low support (56%) for the clade containing *A. aegypti* ADA and *A. gambiae* ADGF-1 due to the almost equally probable topology where *A. gambiae* ADGF-1 and -3 form a monophyletic group (44%). This might indicate that a duplication occurred in the *A. gambiae* lineage after diverging from the other organisms, but this problem will be better resolved once the *A. aegypti* genome is completely sequenced and more ADGF sequences are discovered. The only ambiguity within the six Drosophila genes was found within the clade of ADGF-D gene products. Support for the clade containing *D. melanogaster* ADGF-D and *D. yakuba* ADGF-D was only 72%, and while this topology agrees with the established organismal arrangement, 20% of tree samples found in the run showed *D. melanogaster* ADGF-D basal to the other two Drosophila proteins, and 8% of samples showed *D. yakuba* ADGF-D basal. This ambiguity might be due to the relatively short branch length connecting them, an indication that the three proteins have retained a majority of the same amino acids between them.

Overall, it seems that the protein names given to sequences found in the tblastn searches were correct. For example, all the proteins that were named ADAL are most closely related to each other, without any being placed within other subfamilies. Some protein sequences that were labeled previous to this study and/or published by other groups may in fact be mislabeled. *D. melanogaster* "ADA", *L. longipalpis* "ADA", *C. quinquefasciatus* "ADA", and *A. aegypti* "ADA" do not group with the classic ADA proteins. Instead, *D. melanogaster* "ADA" is a member of the ADAL subfamily, while *L. longipalpis* "ADA", *C. quinquefasciatus* "ADA", and *A. aegypti* "ADA" belong to the insect ADGFs. The mislabeling of these proteins as ADAs was probably due to the fact that each protein was the first adenyl-deaminase member found in that organism, and the sequence similarity to ADA was misleading.

Parsimony analysis

In order to check the accuracy of the ingroup topology produced by Bayesian analysis, a Maximum Parsimony (MP) analysis was performed on the same ingroup alignment. A heuristic search recovered only one most parsimonious tree, however bootstrap support on this tree was not very robust, especially for many internal nodes. In fact, 14 of the 77 Bootstrap values were less than or equal to 50% as placed on the most parsimonious tree (Figure 3-36). Since bootstrap proportions are conservative measures of support, a value of 70% might indicate strong support for a group, compared to the posterior probabilities found in Bayesian analysis (Holder and Lewis, 2003), whereas a support value under 50% would be considered low. The general lack of support for the most parsimonious reconstruction indicated that perhaps not enough tree searches per bootstrap replicate were used in order to find the optimal topology. Alternately, perhaps the use of MP for this data set was not optimal for resolution of internal nodes, since bootstrapping the data did not lend strong support to the infered topology.

Importantly, the subgroups and major clades found in the Bayesian tree were also retained in the MP tree, and the bootstrap support for these clades was high. The bootstrap values were placed on both the MP and Bayesian trees, as shown in Figure 3-36 and Figure 3-35, for comparison. Some internal nodes of the MP tree that were not well supported by bootstrapping were also not well supported with Bayesian posterior probabilities, such as the placement of *U. maydis* and *C. elegans* ADAL within the ADAL subgroup. Also, *E. coli* ADA was placed equally within either the ADE or ADA subgroups for both analyses. Within the ADGF clade, however, there was a major mismatch in the MP tree compared to the Bayesian result. In the MP tree, the *A. californica* MDGF and vertebrate ADGF proteins form a sister group to the insect ADGF-C and -D proteins, with the insect ADGF-A, -B, and -E as the sister to those groups. Also, the ADGF-A2 proteins, instead of being placed as a sister group to the ADGF-A, -B, and -E clade as in the Bayesian analysis, were placed as a sister group to all other vertebrate and insect ADGFs. In effect, the vertebrate proteins were nested within the insect homologues in the MP tree, which produces a major conflict with the Bayesian and established organismal trees. This topology, however, was associated with very low bootstrap values on the most parsimonious MP tree (Figure 3-36).

In order to determine if the MP topology in fact was more probable but had just been overlooked (not sampled) by the Bayesian analysis, MrBayes was run again using the parsimony result as a user defined starting tree. If the MP tree was more highly probable than that previously obtained by MrBayes, then this topology would be expected to persist throughout the run. Instead, the original MrBayes result was again obtained, although the time to stationarity was much less than the other five runs (Figure 3-34), presumably because the starting tree had most of the protein sequences grouped correctly versus a random starting tree. Also, since the support value for the insect ADGF clade in the Bayesian tree shown in Figure 3-35 was 100%, this indicated that there were no instances in which the MP topology that included the ADGF vertebrates was observed, suggesting that the MP topology is not highly probable.

Intron evolution in the ingroup

Preliminary results between the Drosophila *ADGF* and human *CECR1* gene structures showed that two intron positions were shared between species. Due to the wealth of genomic sequence available, the final Bayesian phylogenetic analysis was used to determine the evolution of intron positions among the four ingroup subfamilies. After mapping all intron positions onto the alignment of ingroup proteins, there were a total of 52 distinct intron positions observed in at least one protein sequence. Each intron position was coded into the matrix presented in Figure 3-37, which was then used to reconstruct the most parsimonious intron gain/loss pattern on the inferred Bayesian topology (Figure 3-37, *left*). Only the accelerated transformation reconstruction is shown, in order to more directly test the earliest appearance of each intron. For example, intron location 18 is shown on the figure to be gained in the ancestor of some of the fungal ADGF-1 proteins, and then lost in *M. grisea* ADGF-1. If the changes were delayed instead of accelerated, there would be two instances of intron 18 gain, in both *G. zeae* ADGF-1 and *N. crassa* ADGF-1. Both situations require the same number of steps and are therefore equally parsimonious, but only the accelerated reconstruction is presented. Other introns with an equal number of delayed reconstruction steps include positions 8, 15, 37, and 46.

In general, while none of the bacterial or ADE genes had introns, the fungal and insect genes had between zero and five introns, and the vertebrates had eight to eleven introns, suggesting that organisms that have diverged more recently tend to have more introns. Throughout the entire ingroup, 14 of the 52 intron positions were found in only one taxa. Many of the remaining shared intron positions are found on the branches leading to the vertebrate groups, since five intron positions are only found within the vertebrate ADALs, eight positions within the vertebrate ADAs, and five positions in the vertebrate ADGF homologues (CECR1s). This leaves only 20 intron locations shared among other organisms. Surprisingly, only one intron was found in the exact same place between two different subgroups (position 44; *N. crassa* ADAL and *M. grisea* ADGF-1), but since it was only present in one member of each subgroup, it is most probably a coincidence rather than the persistence of a common intron position in an ancestor of the entire ingroup. Within the ADGF subfamily, positions 19 and 37 are faithfully conserved between the vertebrates and some of the insect subgroups, but this would only suggest that the intron was present in a common ancestor of the metazoans.

Some intron positions between different subgroups are located within just a few base pairs of each other, which might suggest that they have originated in a common ancestor of those subgroups, through the process of intron-sliding (Rogozin et al., 2000). For example, introns 4-7 may have originated from a common ancestor, when intron-sliding is taken into account. These four positions are separated by only nine base pairs total, and are associated with Domain 1 in the ingroup alignment (Figure 3-32). Intron position 4 is located in the vertebrate ADAs just after the conserved "PK" residues, and intron position 5 lies within the G of the "MPKG" motif conserved throughout the vertebrate ADGF homologues. Due to the gap introduced in the ADAs in lieu of this G residue, intron 4 position is actually embedded within the location of intron 5. Intron position 6 is located one codon further, between the vertebrate and *A. gambiae* ADAL "VE" residues, while position 7 is again one codon further, precisely after the "VE" residues, in the fungal ADALs. Even when intron-sliding is taken into account, however, if this intron position did indeed originate in a common ancestor of the four subgroups, the position would need to be lost at least ten different times to account for its absence in all the other ingroup proteins. In this case, it might be more parsimonious to gain the location four times, suggesting that even if intron-sliding is considered, positions 4-7 probably arose independently.

There are three more cases where intron positions might be conserved between different subgroups when intron-sliding is considered. Intron positions 20-22 are separated by only five base pairs in most cases, and involve both the ADA and ADAL vertebrates with the intervening position 21 located in *C. elegans* ADA, but the position would need to be lost at least six times on other branches. Interestingly, intron position 36 found in the vertebrate ADALs is located only four base pairs upstream of position 37, which was described above to be shared between the vertebrate ADGFs and insect ADGF-C & -D clade, but again if this intron was present in a common ancestor of the ingroup, it would need to be lost eight separate times. Finally, position 41 in the ADAL vertebrates is located two base pairs upstream of position 42 found in the vertebrate ADAs, but would need to be lost seven times, which again does not represent the most parsimonious reconstruction.

Within one subgroup, two groups of intron positions may have arisen through intron-sliding. Intron position 29, found in the insect ADGF-C & -D clade, is located just one base pair before position 30, which is found in the insect ADGF-A2, -A, -B, & -E clade (see also Figure 3-30). But in this case, two separate instances of intron gain may be more parsimonious than gain of this intron, sliding to position 30, and loss in the clade containing *A. gambiae* ADGF-1 & -3, which represents three separate steps. The only case in which intron-sliding might be more parsimonious occurs within the vertebrate ADAs, where all eight of the intron positions in this group are precisely conserved, except for an instance of intron-sliding in *D. rerio* ADA (intron 8 slid three bases to position 9). This event represents only two steps (gain plus slide) versus three steps (gain of position 8 in all vertebrates, followed by loss of 8 and gain of 9 only in *D. rerio*), suggesting that intron-sliding is a likely explanation only for this one case. Although it is difficult to decipher how many minor shifts in intron position could be explained by intron-sliding, it seems clear that for this data set, use of this theory is not helpful in suggesting that any of the observed intron positions were present in a common ancestor.

The intron presence/absence binary matrix, excluding taxa without intron data, was used for a MP heuristic search. The resulting tree is presented in Figure 3-38, and shows support for the Bayesian and MP analyses inferred using protein sequences. Except for the placement of single exon genes as polytomies, all of the ADAL and classic ADA members are grouped with their respective subfamily members. Within the ADGF subfamily, the various clades observed in the Bayesian topology depicted in Figure 3-35 were represented, including the vertebrates, Drosophila ADGF-C & -Ds, and Drosophila ADGF-A, -A2, -B & -Es. Interestingly, the MP topology of ADGF vertebrates as a sister to the Drosophila ADGF-C & -Ds was reiterated in this tree, suggesting that these paralogues have diverged less than the other Drosophila ADGFs. The fungal ADGFs, however, did not share many intron positions and were therefore split throughout the tree, underscoring the limitations of this analysis when introns are not conserved. In summary, intron position data is valuable for reconstructing the evolution of intron positions and, in general, supports the evolution of the gene products that contain them.

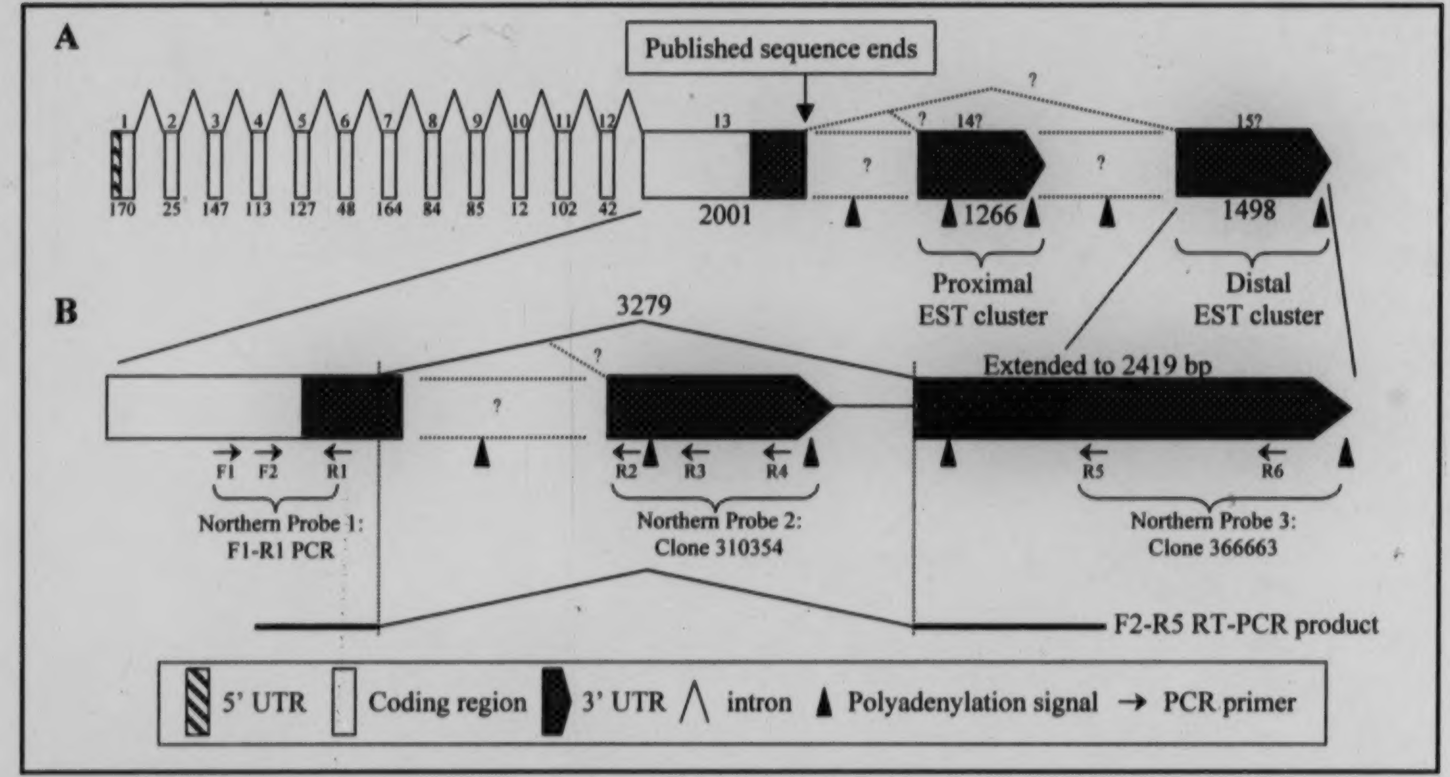
A
Published sequence ends
1
2
3
4
5
6
7
8
9
10
11
12
13
?
? 14?
?
?
15?
170
25
147
113
127
48
164
84
85
12
102
42
2001
1266
1498
Proximal EST cluster
Distal EST cluster
B
3279
?
?
Extended to 2419 bp
F1
F2
R1
R2
R3
R4
R5
R6
Northern Probe 1: F1-R1 PCR
Northern Probe 2: Clone 310354
Northern Probe 3: Clone 366663
F2-R5 RT-PCR product
5' UTR
Coding region
3' UTR
intron
Polyadenylation signal
PCR primer

Figure 3-1. Genetic structure of *IL-17R*. **A**) The published sequence of the gene (accession # U58917) stops in the 3' UTR of exon 13 before the presence of a polyadenylation signal. The exon/intron distances were derived by comparison to PAC 10913. Exons 1-12 are not drawn to scale, but are numbered on top and sizes are given in bp below. **B**) Close-up of the 3' end of *IL-17R* showing PCR primers and Northern probes used to confirm that the two EST clusters are part of the *IL-17R* gene.

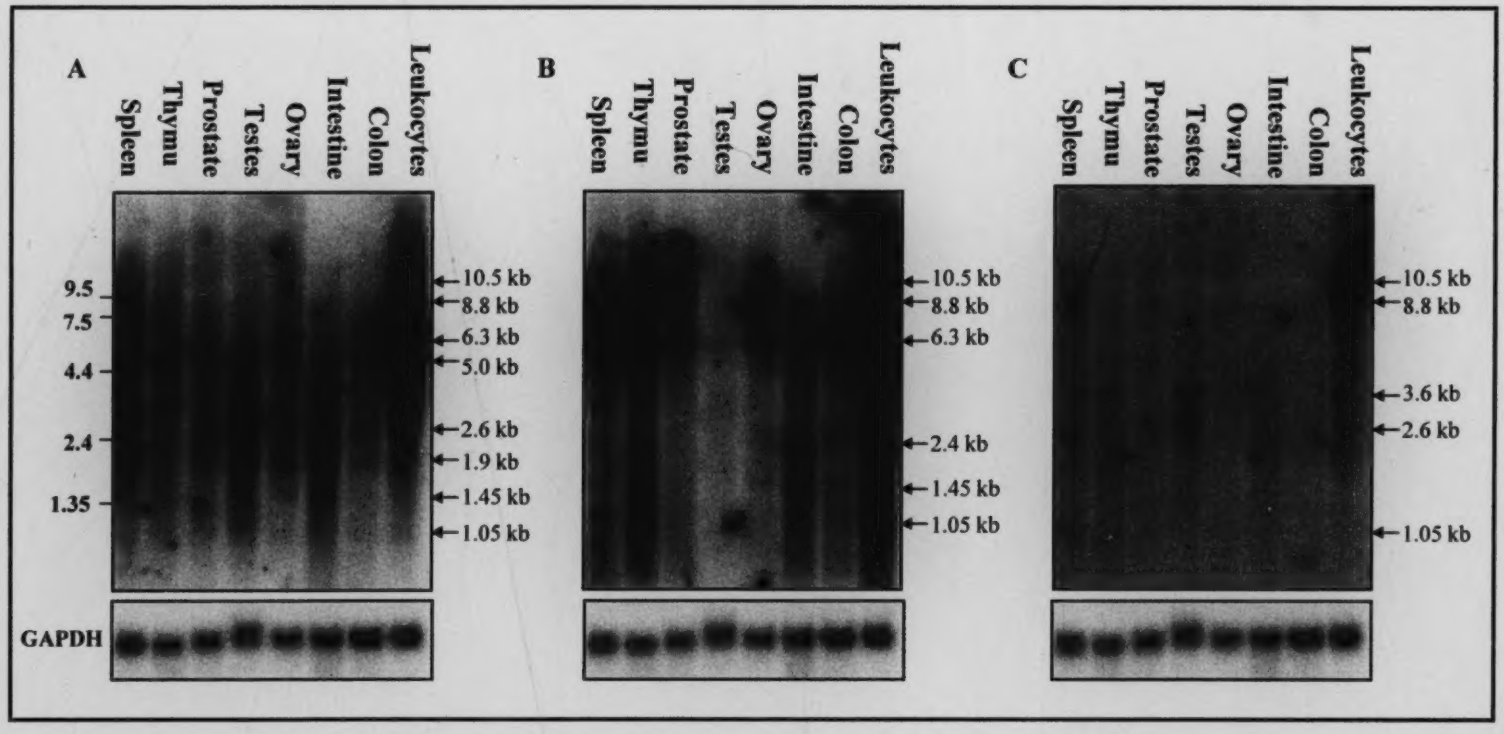
A
Spleen
Thymu
Prostate
Testes
Ovary
Intestine
Colon
Leukocytes
9.5
7.5
4.4
2.4
1.35
10.5 kb
8.8 kb
6.3 kb
5.0 kb
2.6 kb
1.9 kb
1.45 kb
1.05 kb
GAPDH
B
Spleen
Thymu
Prostate
Testes
Ovary
Intestine
Colon
Leukocytes
10.5 kb
8.8 kb
6.3 kb
2.4 kb
1.45 kb
1.05 kb
C
Spleen
Thymu
Prostate
Testes
Ovary
Intestine
Colon
Leukocytes
10.5 kb
8.8 kb
3.6 kb
2.6 kb
1.05 kb

Figure 3-2. Expression analysis of human *IL-17R*. An adult Clontech Northern blot (lot#7100045) was probed sequentially with **A)** a PCR probe made with primers IL-F1 and IL-R1, **B)** EST 310354, which represents the proximal EST cluster, and C) EST 366663, which represents the distal EST cluster. Multiple bands are detectable for each probe as indicated on the right edge of each Northern blot. The GAPDH loading control is beneath each blot.

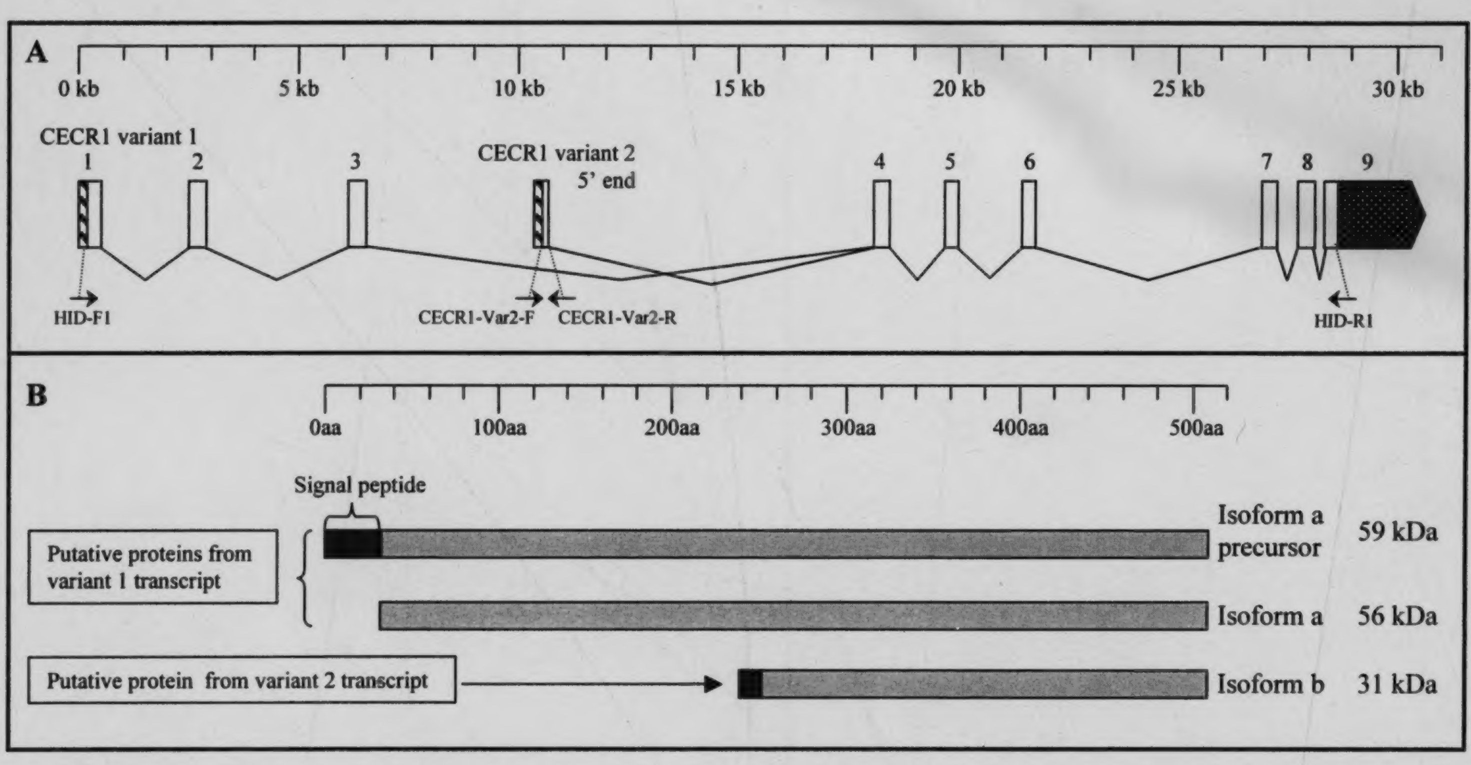
A
0 kb
5 kb
10 kb
15 kb
20 kb
25 kb
30 kb
CECR1 variant 1
1
2
3
CECR1 variant 2
5' end
4
5
6
7
8
9
HID-F1
CECR1-Var2-F
CECR1-Var2-R
HID-R1
B
0aa
100aa
200aa
300aa
400aa
500aa
Signal peptide
Putative proteins from variant 1 transcript
Isoform a precursor
59 kDa
Isoform a
56 kDa
Putative protein from variant 2 transcript
Isoform b
31 kDa

Figure 3-3. Genomic structure and predicted protein sequences of human *CECR1*. **A)** The *CECR1* gene spans 30,582 bp of genomic sequence. The *CECR1* cDNA (also referred to as *CECR1v1*) contains 9 exons comprising 3923 bp. *CECR1v2* starts in intron 3 of *CECR1v1*; its cDNA encompasses 3071 bp. The primers used to make Northern probes are shown. **B)** Hypothetical protein sequences derived from the *CECR1* locus. The CECR1v1 full length sequence (isoform a precursor) is 511 aa in length, and 59 kDa in size. Isoform a (without signal peptide) has a predicted weight of 56 kDa. The putative CECR1v2 product (isoform b) is 31 kDa.

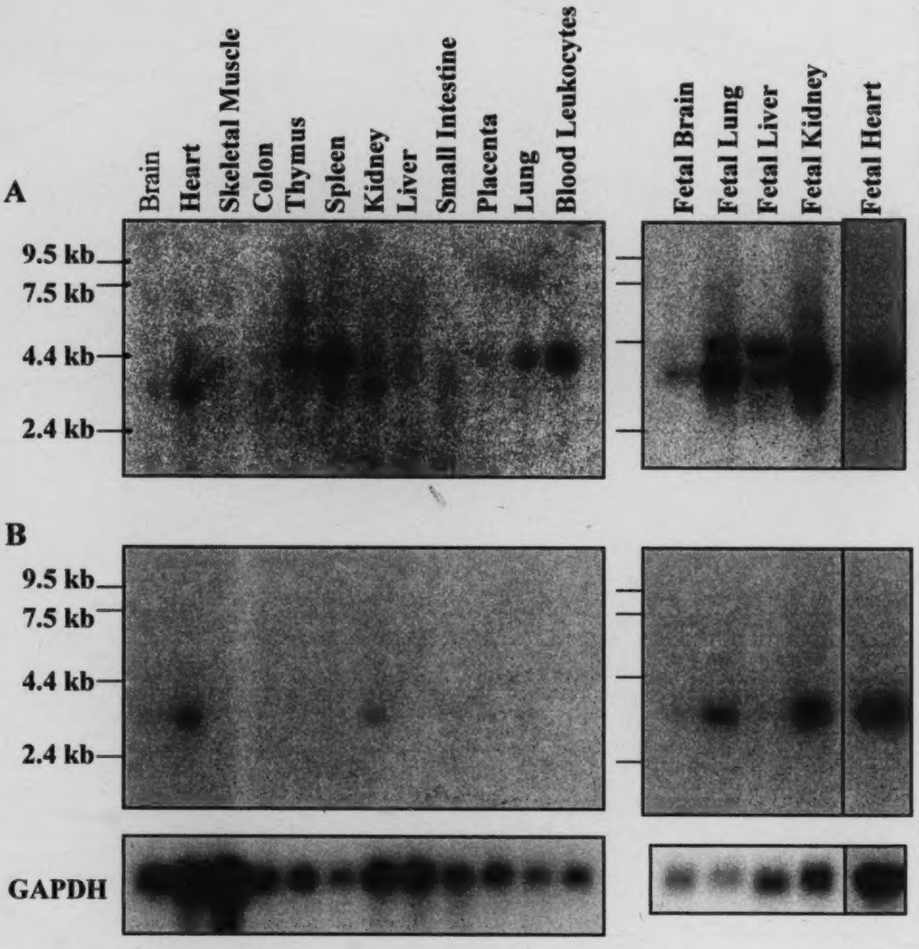
A
Brain
Heart
Skeletal Muscle
Colon
Thymus
Spleen
Kidney
Liver
Small Intestine
Placenta
Lung
Blood Leukocytes
Fetal Brain
Fetal Lung
Fetal Liver
Fetal Kidney
Fetal Heart
9.5 kb
7.5 kb
4.4 kb
2.4 kb
B
9.5 kb
7.5 kb
4.4 kb
2.4 kb
GAPDH

Figure 3-4. Expression analysis of the different *CECR*1 isoforms. Human adult (lot # 8030922) and fetal (lot # 7100219) Clontech blots plus fetal blot NF2 (Invitrogen, lot # 7910041); only the heart lane is shown) were used. Blots were probed with either **A)** a PCR probe made with primers HID-F1& -R1 which includes the entire *CECR1v1* coding region, or **B)** a variant 2 specific PCR probe made with primers CECR1-Var2-F & -R. Bands of 4.4 or 3.5 kb are observed in panel A, but only the 3.5 kb band is detected in panel B. The GAPDH loading control is shown beneath each blot.

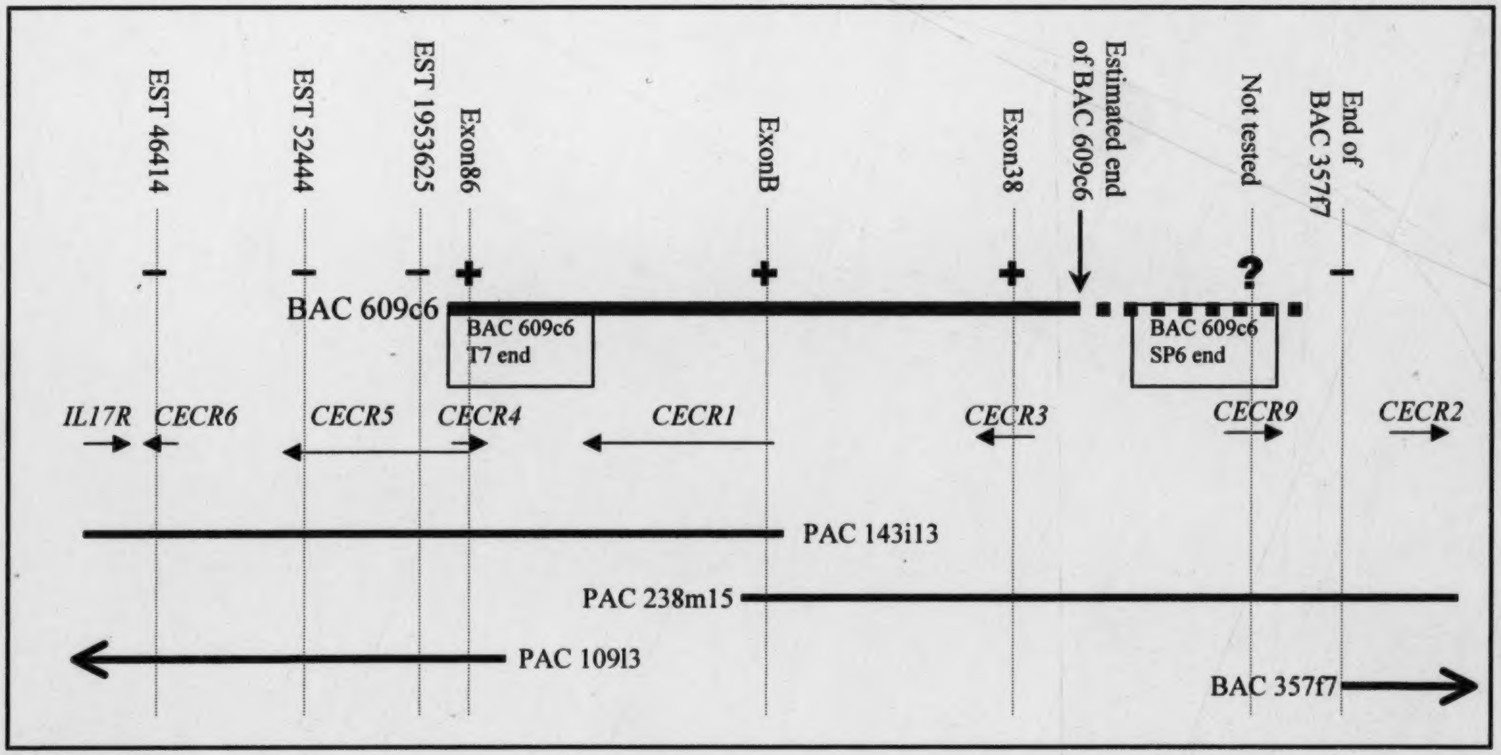
EST 46414
EST 52444
EST 1953625
Exon86
ExonB
Exon38
Estimated end of BAC 609c6
Not tested
End of BAC 357f7
−
−
−
+
+
+
?
−
BAC 609c6
BAC 609c6 T7 end
BAC 609c6 SP6 end
IL17R
CECR6
CECR5
CECR4
CECR1
CECR3
CECR9
CECR2
PAC 143i13
PAC 238m15
PAC 109I3
BAC 357f7

Figure 3-5. Characterization of human BAC 609c6. Southern analysis of BAC 609c6 (blue line) was performed using various EST and exon probes listed at the top of the figure, and whether each probe gave a positive (+) or negative (-) signal for BAC 609c6 is indicated. A probe for CECR9 was not tested, but the estimated length of BAC 609c6 suggests that it is not present (marked as a ? here). The end clone of BAC 357f7 was previously published (Johnson et al., 1999). PACs 109i3, 143i13, and 238m15, and BAC 357f7 are shown as thick lines, and genes are shown as thin arrows, for comparison of probe locations.

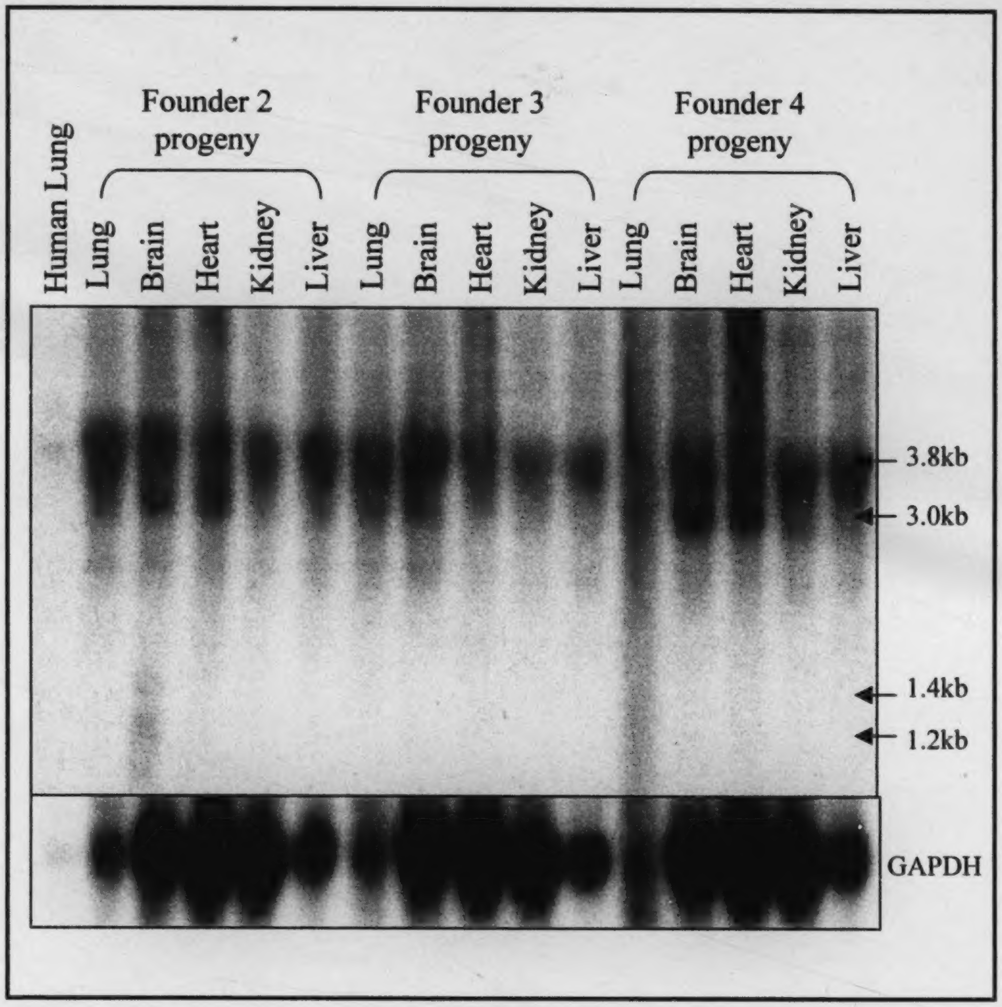
Human Lung
Founder 2 progeny
Founder 3 progeny
Founder 4 progeny
Lung
Brain
Heart
Kidney
Liver
Lung
Brain
Heart
Kidney
Liver
Lung
Brain
Heart
Kidney
Liver
3.8kb
3.0kb
1.4kb
1.2kb
GAPDH

Figure 3-6. Expression analysis of human *CECR1* in transgenic mice. A Northern blot of transgenic mouse RNA was probed with the human *CECR1* coding region (HID-F1/R1) PCR probe. Human Lung (lane 1) acts as a control and size comparison. Bands of 3.8 and 3.0 kb can be observed in most transgenic mouse tissues, while the two smaller bands are unique to brain tissue of Founder 2 progeny. The GAPDH loading control is shown below the blot.

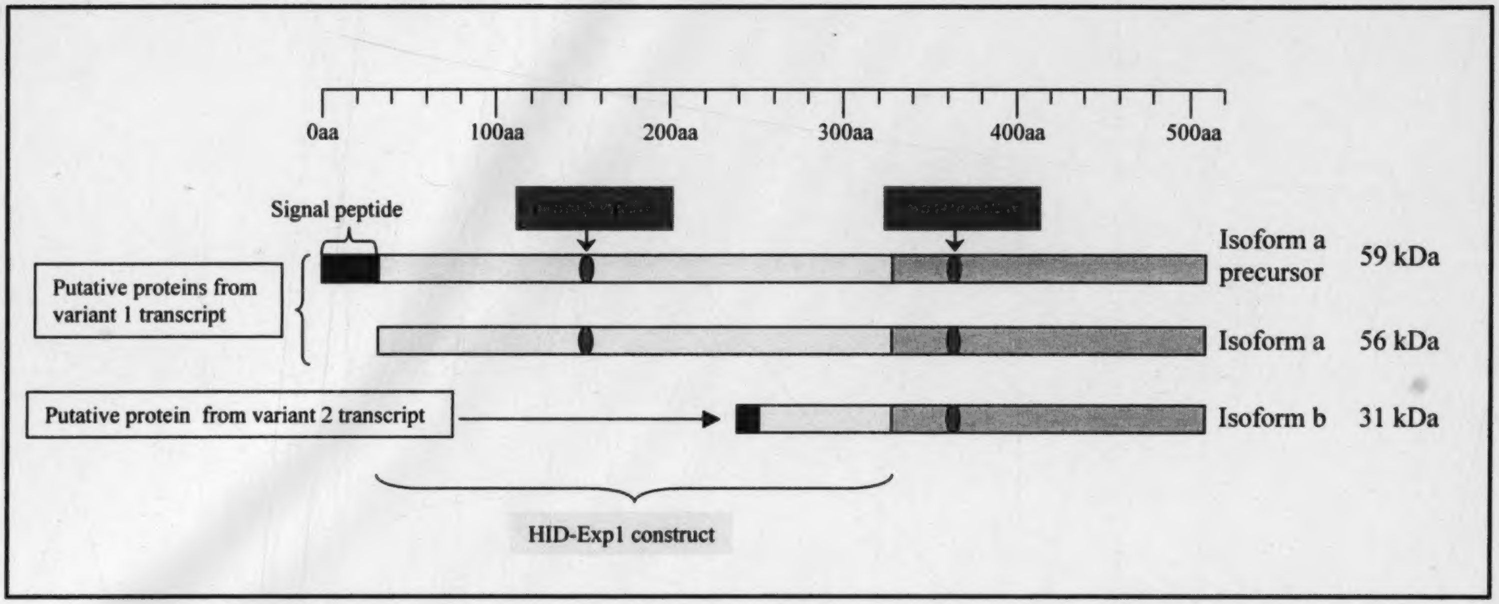
0aa
100aa
200aa
300aa
400aa
500aa
Signal peptide
Putative proteins from variant 1 transcript
Putative protein from variant 2 transcript
Isoform a precursor
59 kDa
Isoform a
56 kDa
Isoform b
31 kDa
HID-Exp1 construct

Figure 3-7. Hypothetical CECR1 protein sequences and locations of antigens. The full length product (isoform a precursor) is 511 aa in length, and 59 kDa in size. Isoform a (without signal peptide) has a predicted weight of 56 kDa. The putative CECR1v2 product (isoform b) is 31 kDa. The location of the expression construct (HID-Exp1) and two peptides (CECR1-Pep1 and -Pep2) used to make CECR1 antibodies is depicted on all three putative gene products.

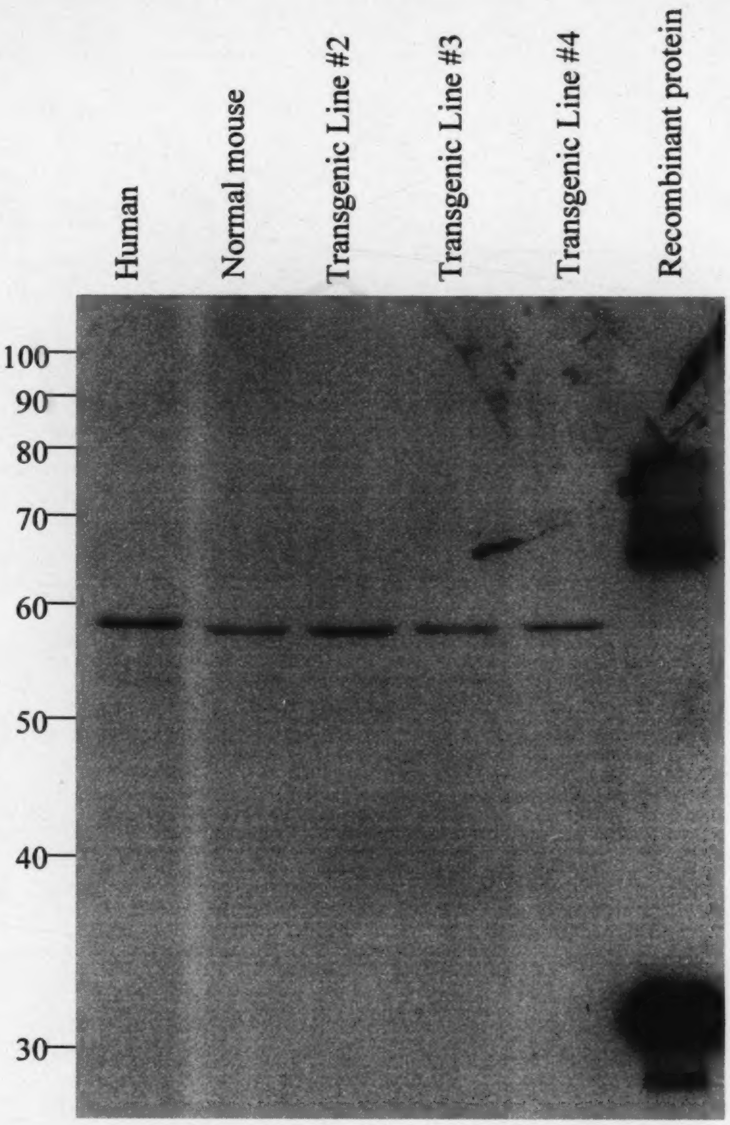
Human
Normal mouse
Transgenic Line #2
Transgenic Line #3
Transgenic Line #4
Recombinant protein
100
90
80
70
60
50
40
30

Figure 3-8. Western analysis of spleen extracts using the 0A1 antibody. Spleen protein lysates were extracted from the indicated sources. Band sizes indicated on the left are in kDa. The recombinant protein (HIDExp1) is the positive control. Note the presence of a cross-reacting ~58 kDa band in the normal mouse sample.

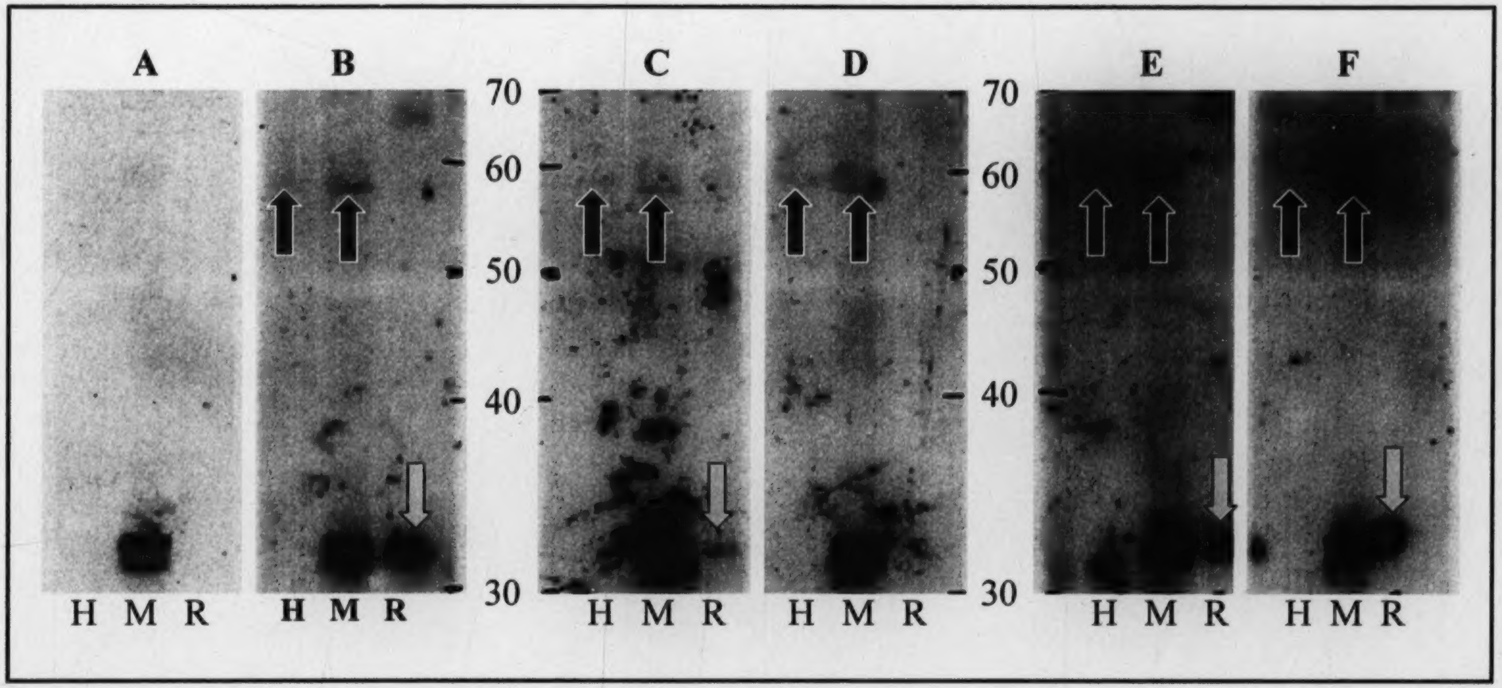

Figure 3-9. Competition assay using the HIDExp1 recombinant protein. Membrane strips containing the same protein samples were probed with **A)** 0A1 preimmune serum, **B)** 0A1 final bleed serum, or 0A1 final bleed serum competed with **C)** 10 times or **D)** 50 times the amount of HIDExp1 recombinant protein, or **E)** 10 times or **F)** 50 times the amount of BSA as a negative control. Blue (dark) arrows indicate the human and mouse bands, while yellow (light) arrows show the HIDExp1 band. Numbers indicate the molecular weight, in kDa. H, human spleen lysate; M, mouse spleen lysate; R, recombinant HIDExp1 protein. There are two spurious ~32 kDa bands in the mouse lane.

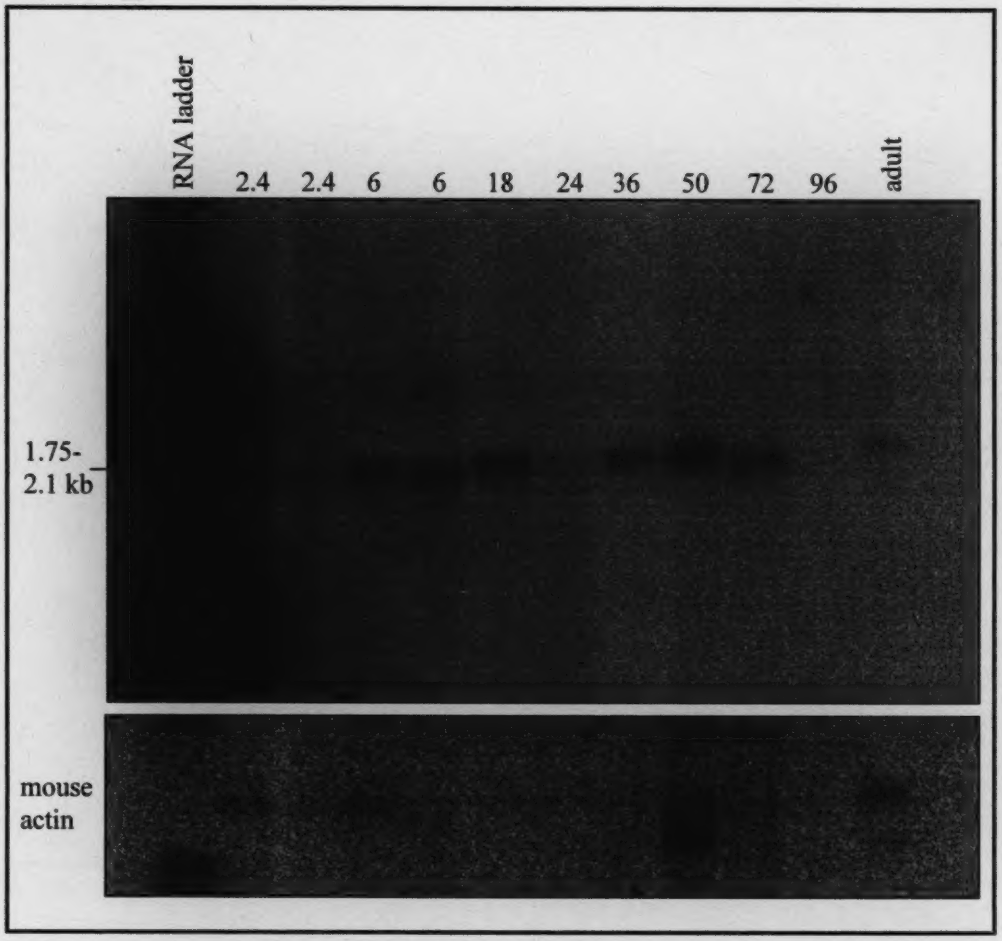

Figure 3-10. Expression analysis of zebrafish *CECR1-1*. The blot was provided by Jon Staav and comprises RNA isolated by various students from Jon's class. Numbers refer to developmental stages, in hours post fertilization (hpf). The blot was probed by the author, using a ssDNA probe. The band size varies from 1.75 to 2.1 kb, depending on the stage, which may be due to differences in loading the samples by the students. A mouse β-actin probe was used as the loading control, since no zebrafish control could be obtained, and is shown in the bottom panel.

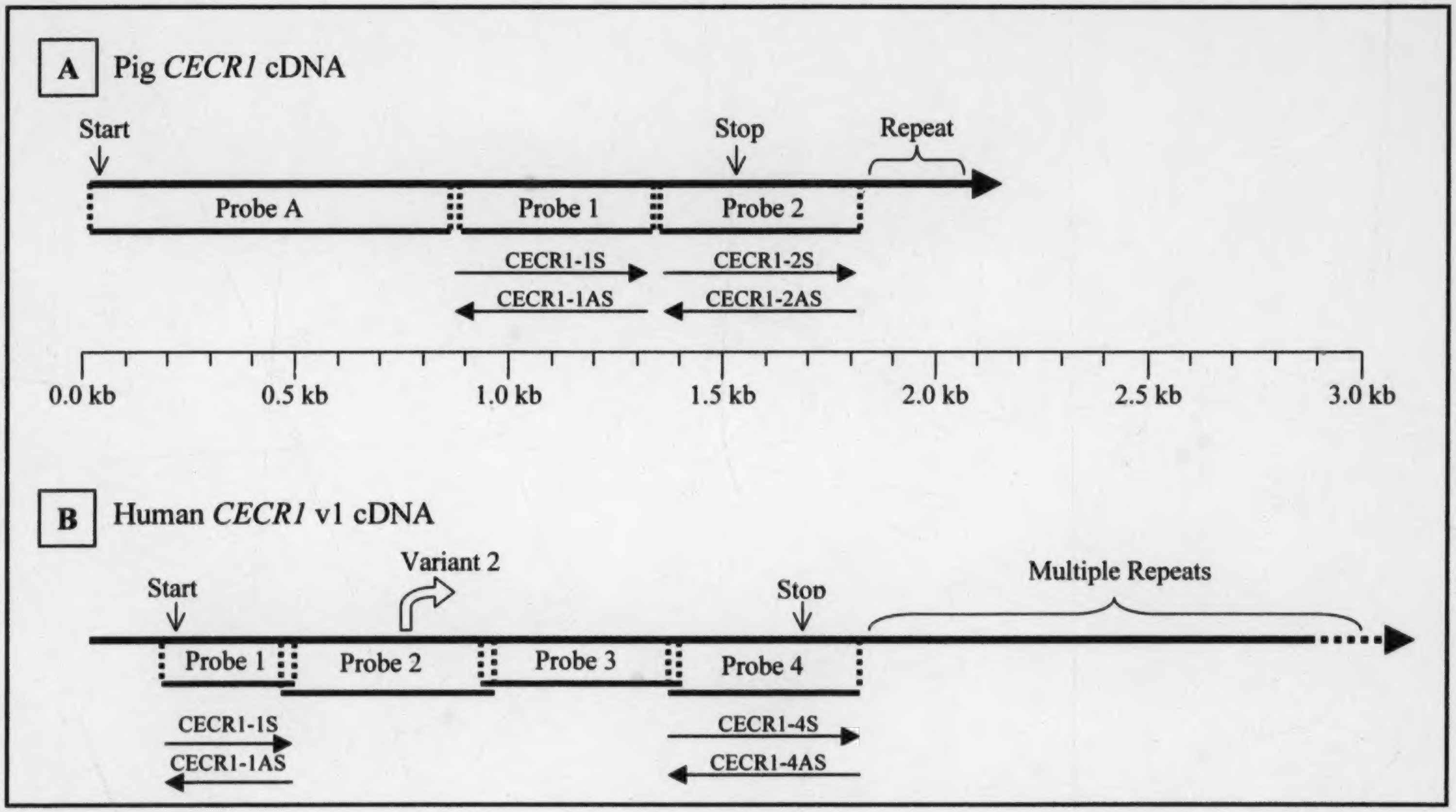
A
Pig CECR1 cDNA
Start
Stop
Repeat
Probe A
Probe 1
Probe 2
CECR1-1S
CECR1-1AS
CECR1-2S
CECR1-2AS
0.0 kb
0.5 kb
1.0 kb
1.5 kb
2.0 kb
2.5 kb
3.0 kb
B
Human CECR1 v1 cDNA
Variant 2
Start
Stop
Multiple Repeats
Probe 1
Probe 2
Probe 3
Probe 4
CECR1-1S
CECR1-1AS
CECR1-4S
CECR1-4AS

Figure 3-11. Locations of *in situ* hybridization probes in the **A)** pig and **B)** human *CECR1* genes. Each cDNA is represented as an arrow, from 5' to 3', with the probes indicated underneath. Start and stop refer to translation. *CECR1v2* shares the human sequence starting within probe 2.

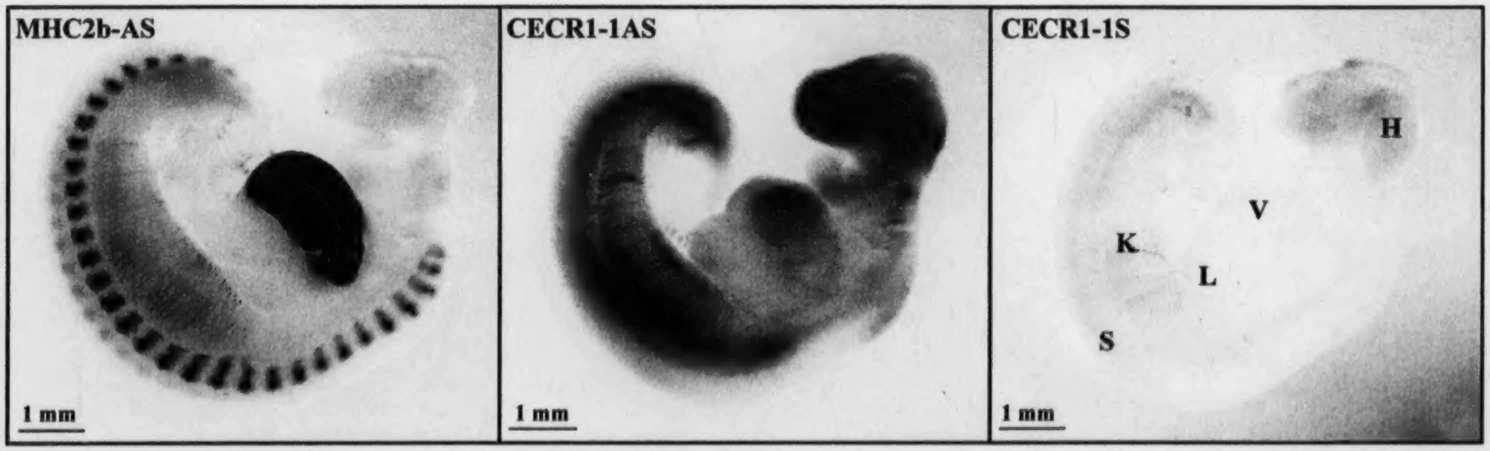

Figure 3-12. Whole-mount *in situ* hybridization of day 20 pig embryos using various probes. MHC2b was used as a positive control, which stains the heart atria and ventricles, the myotomes along the spine, and the kidney tubules. The CECR1-1AS probe shows staining in the heart, kidney tubules and spine, with less staining in the liver. The CECR1-1S probe shows some non-specific trapping of the probe in the head. H, head; K, mesonephric kidney; L, liver; S, spine; V, heart ventricle. Probe staining is depicted as a purple colour.

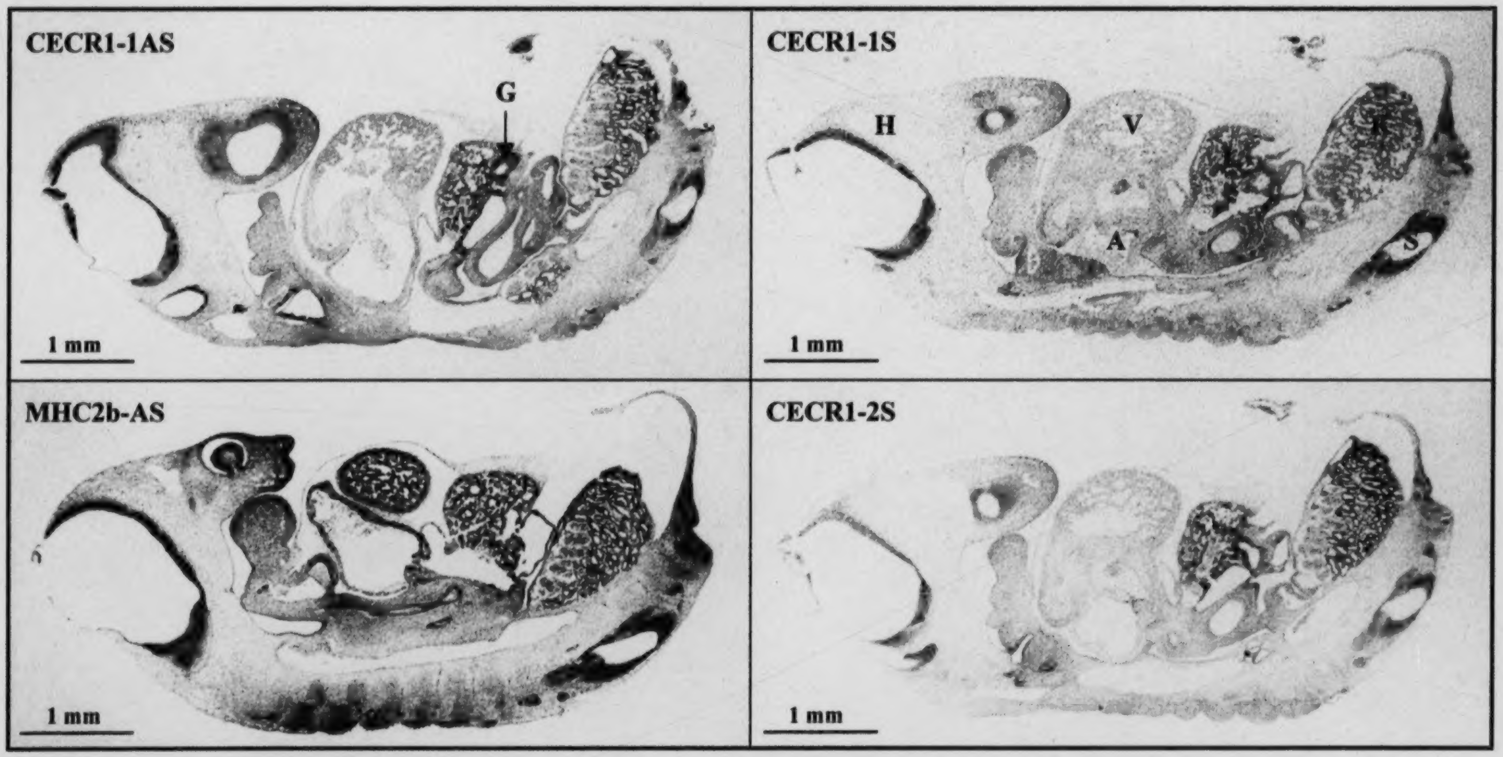
CECR1-1AS
G
1 mm
CECR1-1S
H
V
A
S
1 mm
MHC2b-AS
1 mm
CECR1-2S
1 mm

Figure 3-13. *In situ* hybridization of day 20 pig embryo sections using various probes. MHC2b was used as a positive control. For each embryo section, the head appears on the left, with dorsal side down. A, heart atria; G, gut epithelium; H, head; K, mesonephric kidney; L, liver; S, spinal chord; V, heart ventricle. Probe staining is purple/red, while the counter-stain is green/blue.

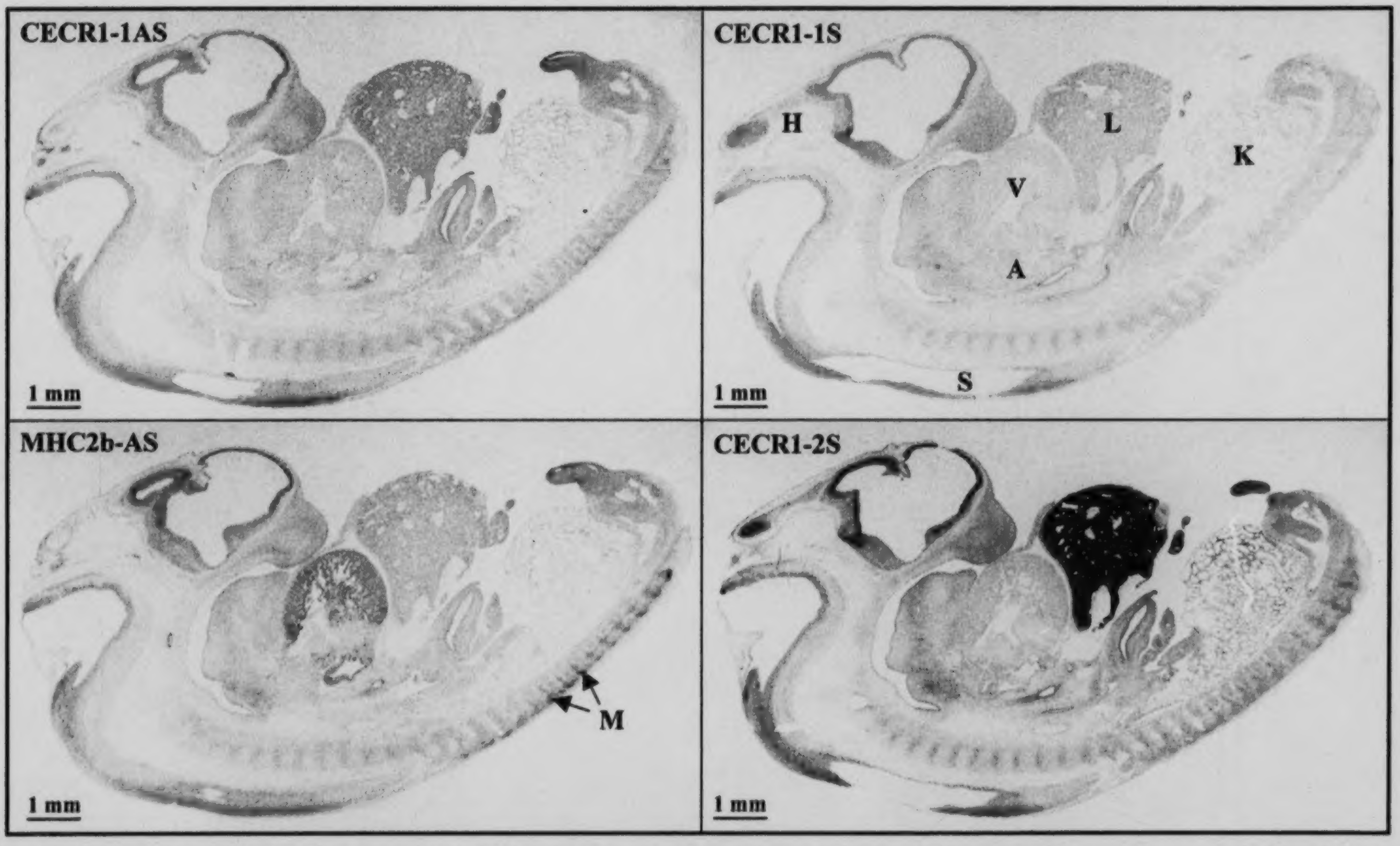
CECR1-1AS
1 mm
CECR1-1S
H
L
K
V
A
S
1 mm
MHC2b-AS
M
1 mm
CECR1-2S
1 mm

Figure 3-14. *In situ* hybridization of day 28 pig embryo sections using various probes. MHC2b acted as a positive control. For each embryo section, the head appears on the left, with dorsal aspect down. A, heart atria; H, head; K, mesonephric kidney; L, liver; M, myotome; S, spinal chord; V, heart ventricle. Probe staining is purple/red; counter-stain is green/blue.

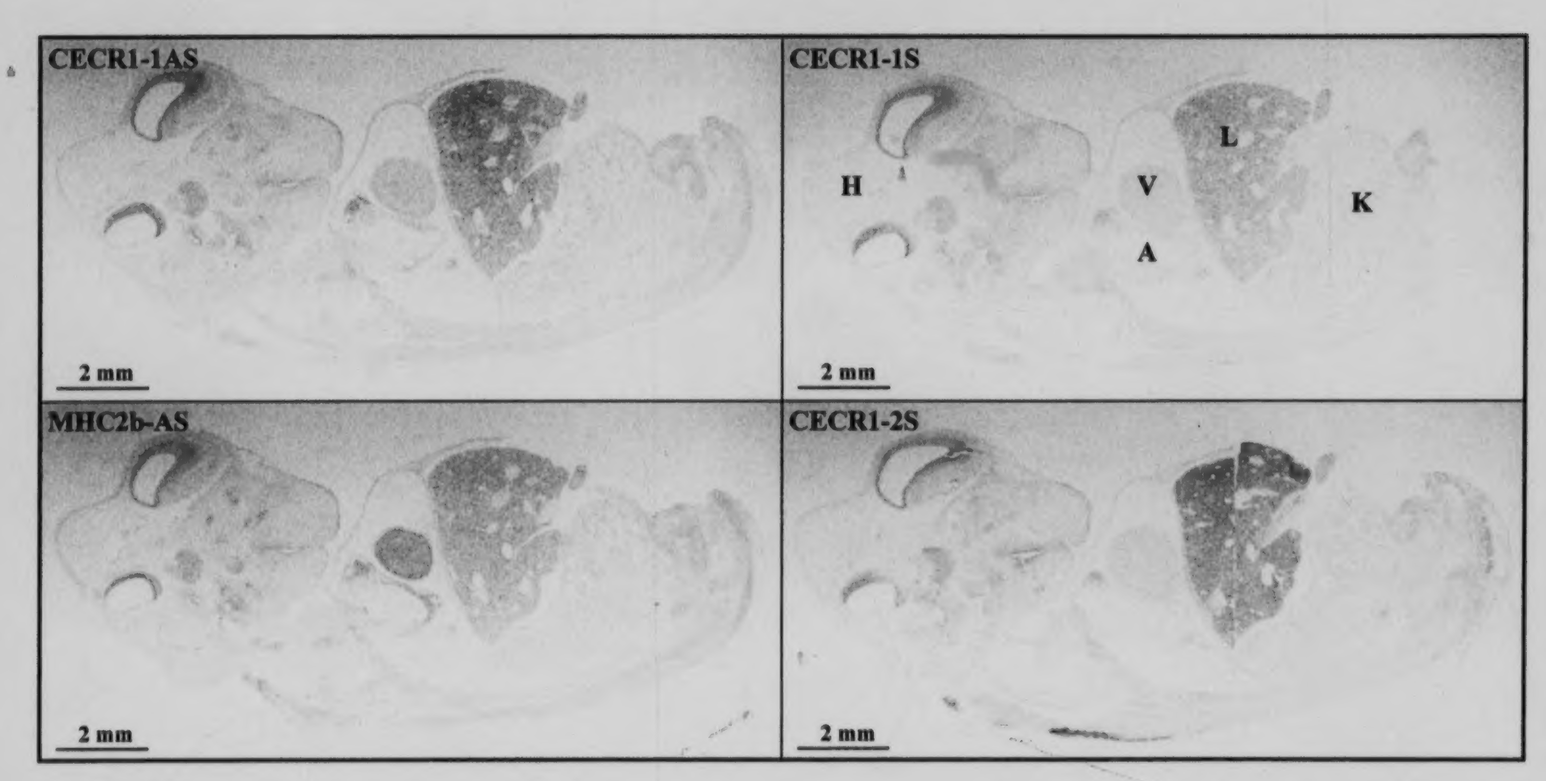
CECR1-1AS
2 mm
CECR1-1S
H
V
A
L
K
2 mm
MHC2b-AS
2 mm
CECR1-2S
2 mm

Figure 3-15. *In situ* hybridization of day 31 pig embryo sections using various probes. MHC2b was used as a positive control. For each embryo section, the head appears on the left, with dorsal aspect down. A, heart atria; H, head; K, mesonephric kidney; L, liver; V, heart ventricle. Probe staining is purple/red, while the counter-stain is green/blue.

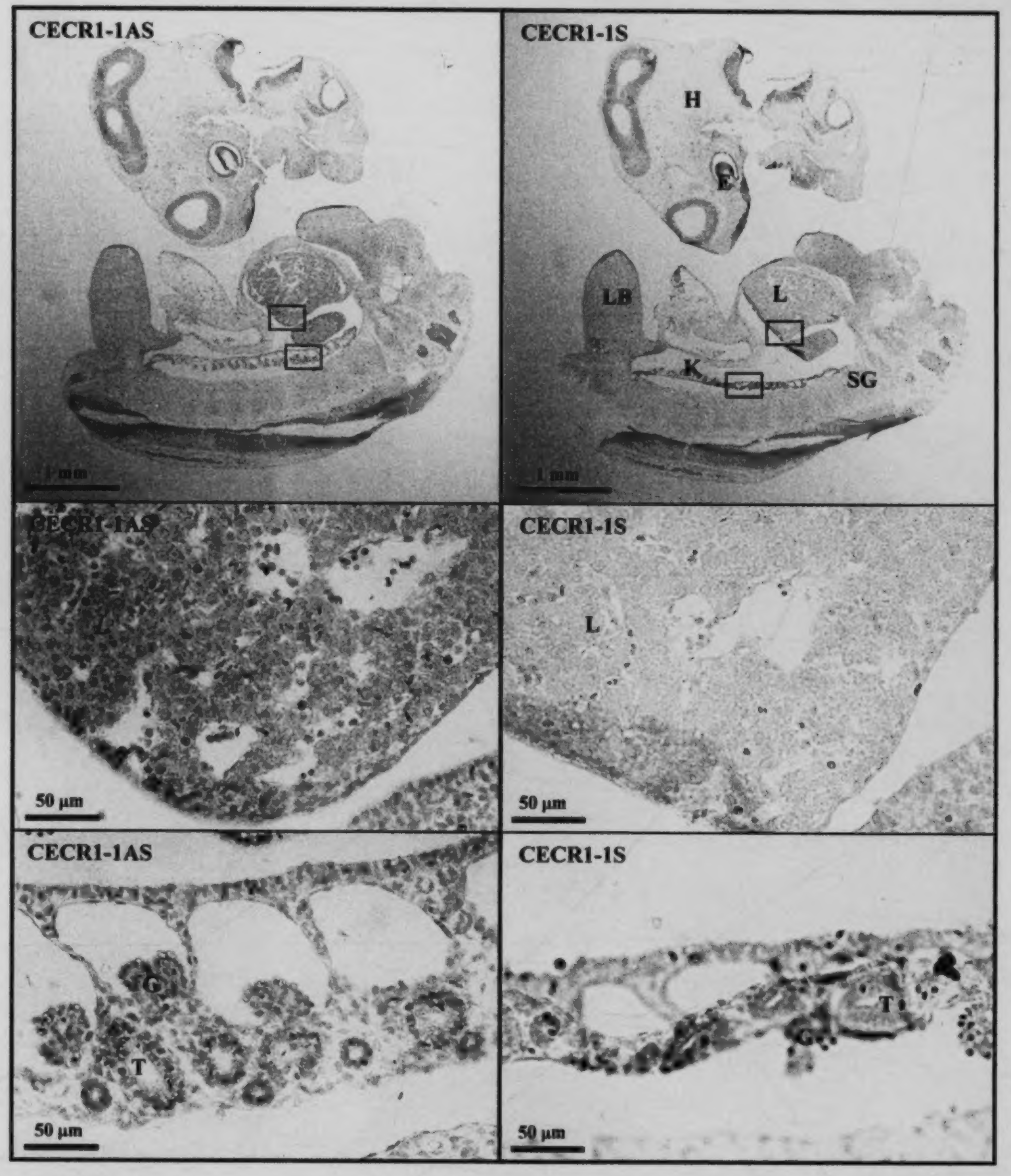
CECR1-1AS
CECR1-1S
H
E
LB
L
K
SG
1 mm
1 mm
CECR1-1AS
CECR1-1S
L
50 µm
50 µm
CECR1-1AS
CECR1-1S
G
T
T
G
50 µm
50 µm

Figure 3-16. RNA *in situ* hybridization of human fetal day 34 sections using either the (*left*) CECR1-1AS or (*right*) CECR1-1S probe. (*Top*) Serial sections from the same embryo. (*Bottom*) Close-up views of the boxes shown in the embryos above, highlighting the liver and kidney staining. E, eye; G, kidney glomerulus; H, head; K, mesonephric kidney; L, liver; LB, limb bud; SG, spinal ganglia; T, kidney tubule. Probe staining is purple/red, while the counter-stain is green/blue.

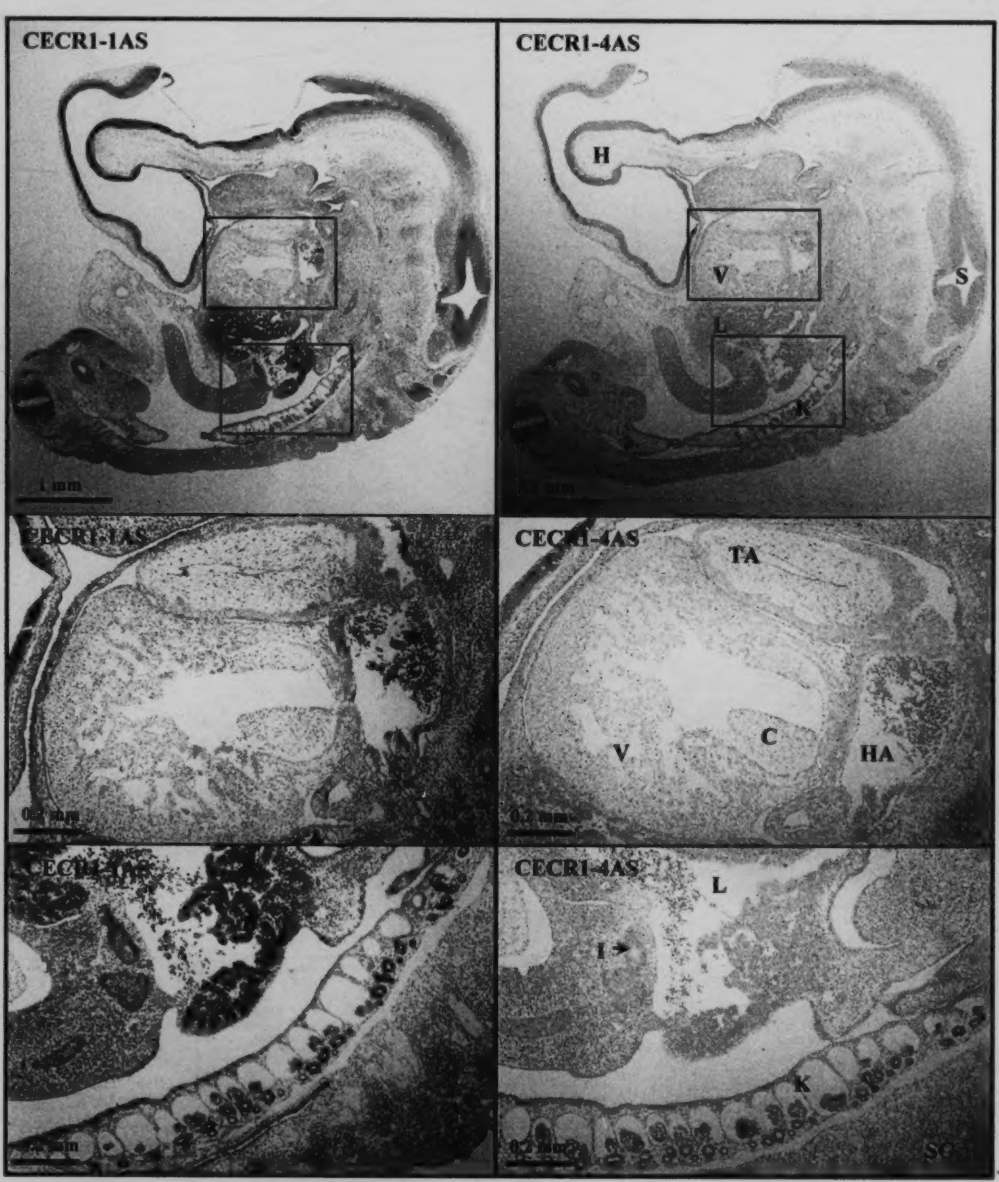
CECR1-1AS
1 mm
CECR1-4AS
H
V
S
L
K
1 mm
CECR1-1AS
0.2 mm
CECR1-4AS
TA
V
C
HA
0.2 mm
CECR1-1AS
0.2 mm
CECR1-4AS
L
I
K
SG
0.2 mm

Figure 3-17. RNA *in situ* hybridization of human fetal day 34 sections with the (*left*) CECR1-1AS or (*right*) CECR1-4AS probe. (*Top*) Serial sections from the same embryo as the previous figure, but in a more medial plane. (*Bottom*) Close-ups of boxes in the embryos above. C, endocardial cushions; H, head; HA, heart atrium; I, small intestine; K, mesonephric kidney; L, liver; S, spinal chord; SG, spinal ganglia; TA, truncus arteriosus; V, heart ventricle. Probe staining is purple/red; counter-stain is green/blue.

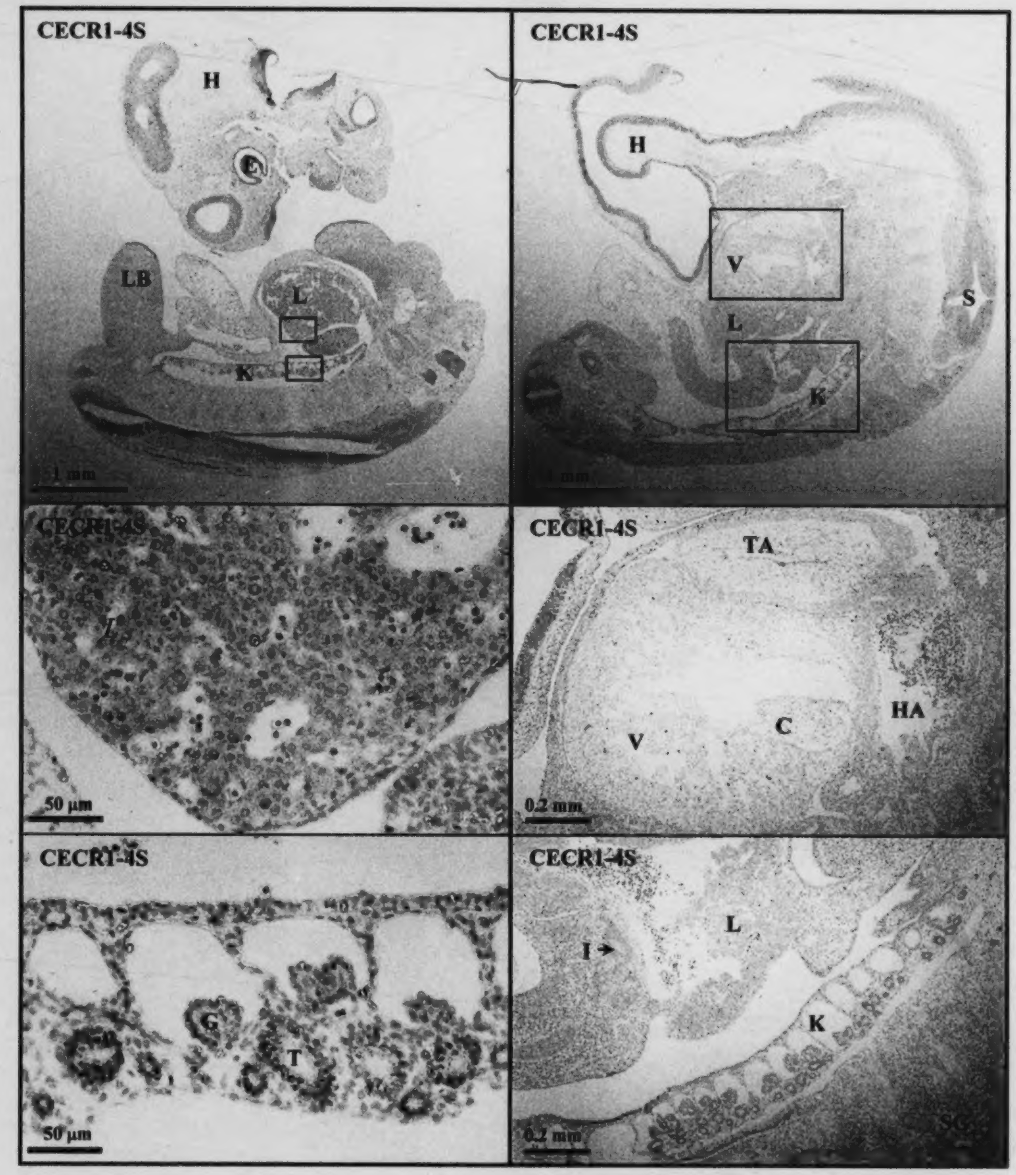
CECR1-4S
H
E
LB
L
K
1 mm
CECR1-4S
H
V
L
S
K
1 mm
CECR1-4S
L
50 µm
CECR1-4S
TA
HA
C
V
0.2 mm
CECR1-4S
G
T
50 µm
CECR1-4S
L
I
K
SG
0.2 mm

Figure 3-18. RNA *in situ* hybridization of human fetal day 34 sections using the CECR1-4S probe. (*Top*) Sections from different planes of the same embryo in the previous two figures. (*Bottom*) Close-ups of boxes in the embryos above. C, endocardial cushions; E, eye; G, glomerulus; H, head; HA, heart atrium; I, small intestine; K, mesonephric kidney; L, liver; LB, limb bud; S, spinal chord; SG, spinal ganglia; T, kidney tubule; TA, truncus arteriosus; V, heart ventricle. Probe staining is purple/red; counter-stain is green/blue.

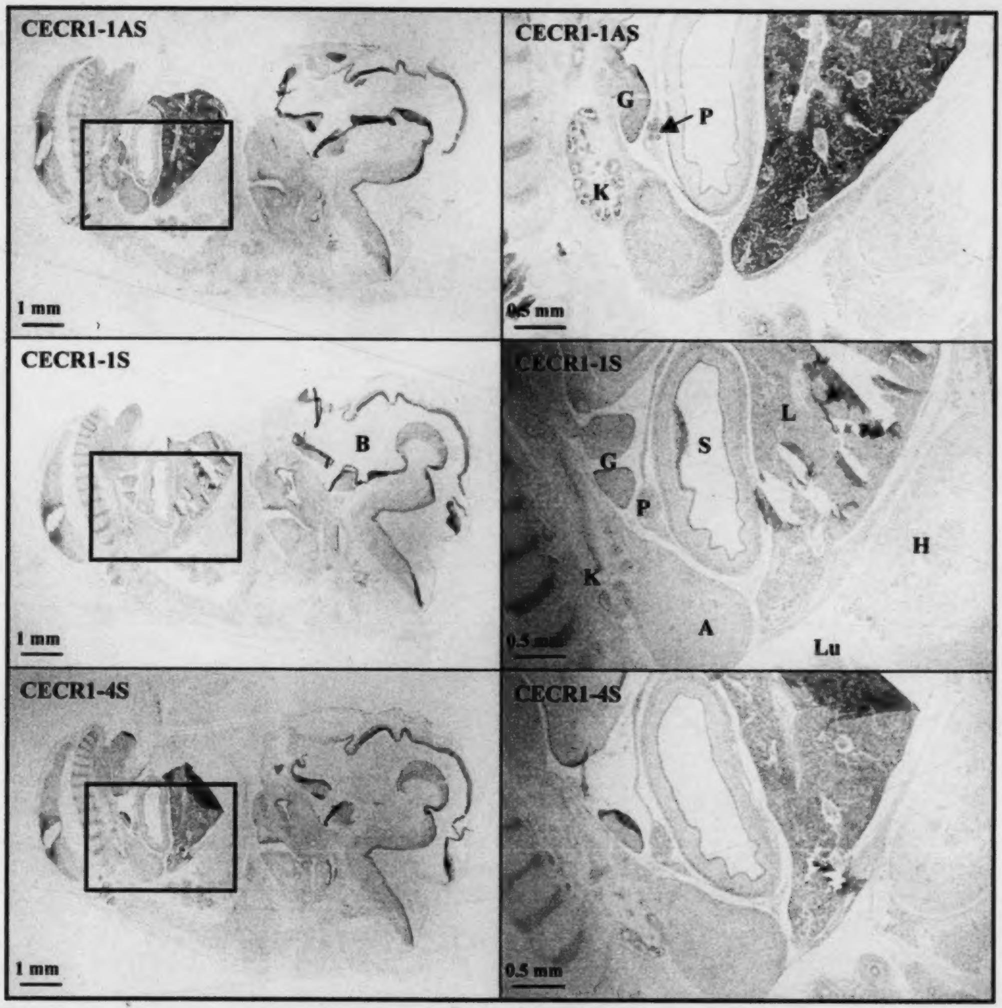
CECR1-1AS
1 mm
CECR1-1AS
G
P
K
0.5 mm
CECR1-1S
B
1 mm
CECR1-1S
L
S
G
P
H
K
A
Lu
0.5 mm
CECR1-4S
1 mm
CECR1-4S
0.5 mm

Figure 3-19. RNA *in situ* hybridization of human fetal day 47 sections using various CECR1 probes. (*Left*) Three adjacent sections from the same embryo. The head appears on the right, with dorsal aspect down. (*Right*) Close-up views of the boxes shown in embryos on the left. A, adrenal gland; B, brain; G, gonad; H, heart; K, metanephric kidney; L, liver; Lu, lung; P, pancreas; S, stomach. Probe staining is purple/red, while the counter-stain is green/blue.

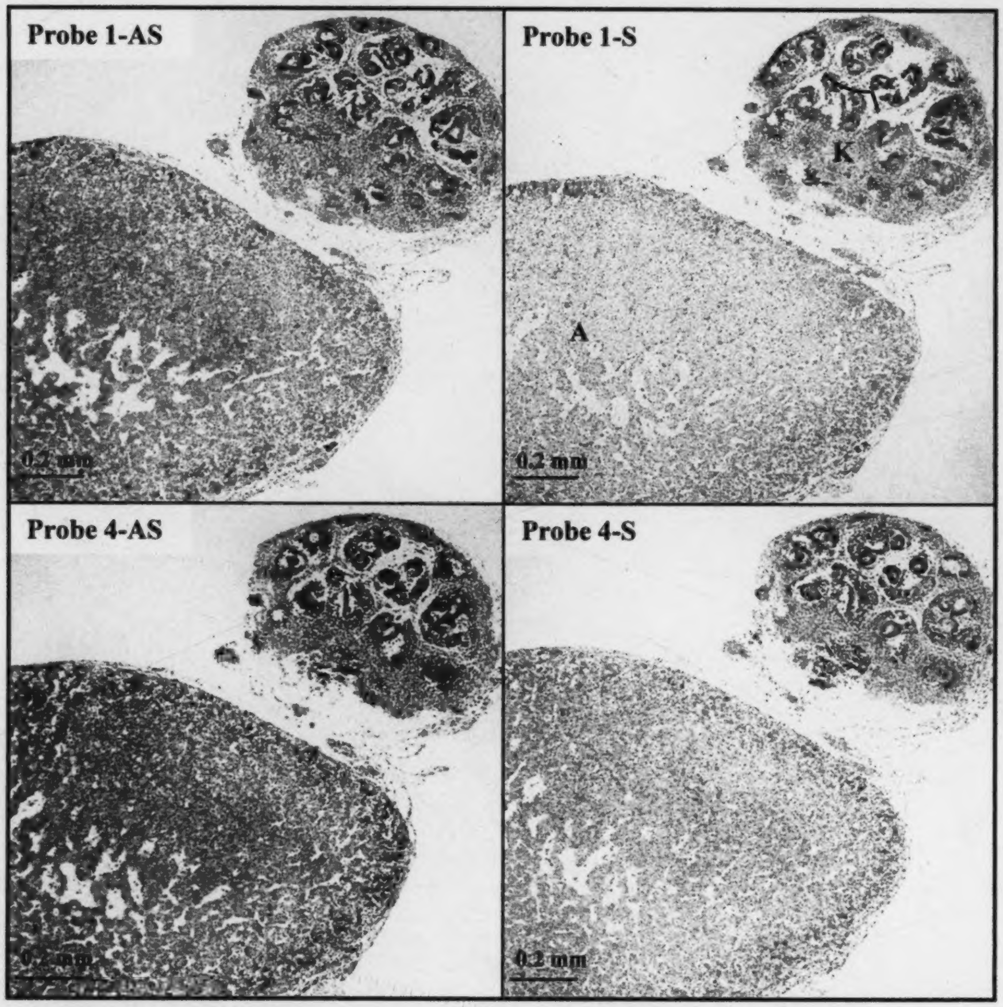
Probe 1-AS
Probe 1-S
K
A
Probe 4-AS
Probe 4-S
0.2 mm
0.2 mm
0.2 mm
0.2 mm

Figure 3-20. RNA *in situ* hybridization of human fetal week 8.5 sections showing the adrenal gland (A) and metanephric kidney (K). Staining due to the various probes is indicated as a purple colour, while the counter-stain is green. Note the staining in the adrenal gland when the 1-AS, 4-AS, and 4-S probes are used, while the 1-S probe section is negative.

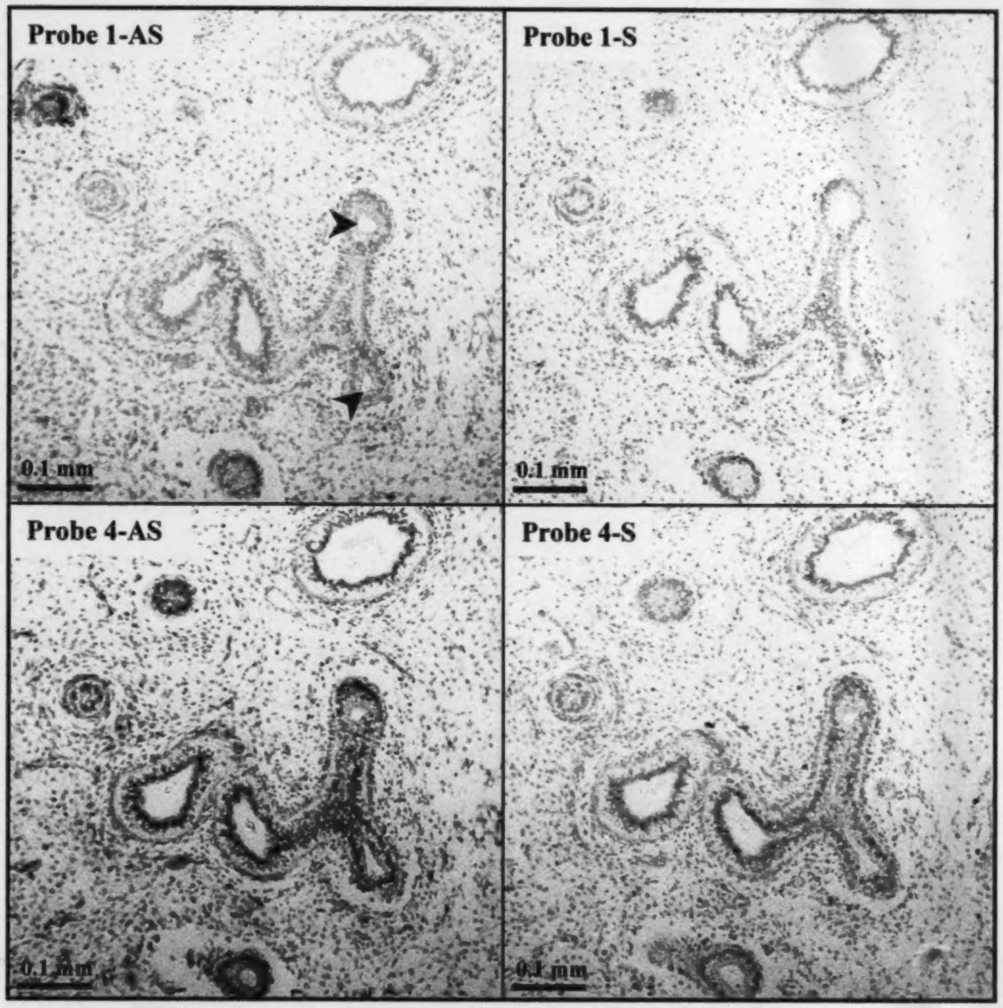
Probe 1-AS
Probe 1-S
Probe 4-AS
Probe 4-S
0.1 mm
0.1 mm
0.1 mm
0.1 mm

Figure 3-21. RNA *in situ* hybridization of human fetal week 8.5 lung sections. Staining due to the various probes is indicated as a purple colour, while the counter-stain is green. Note the staining of the epithelial lining (arrowheads) and surrounding mesoderm when the 1-AS, 4-AS, and 4-S probes are used, while the section is negative with the 1-S probe.

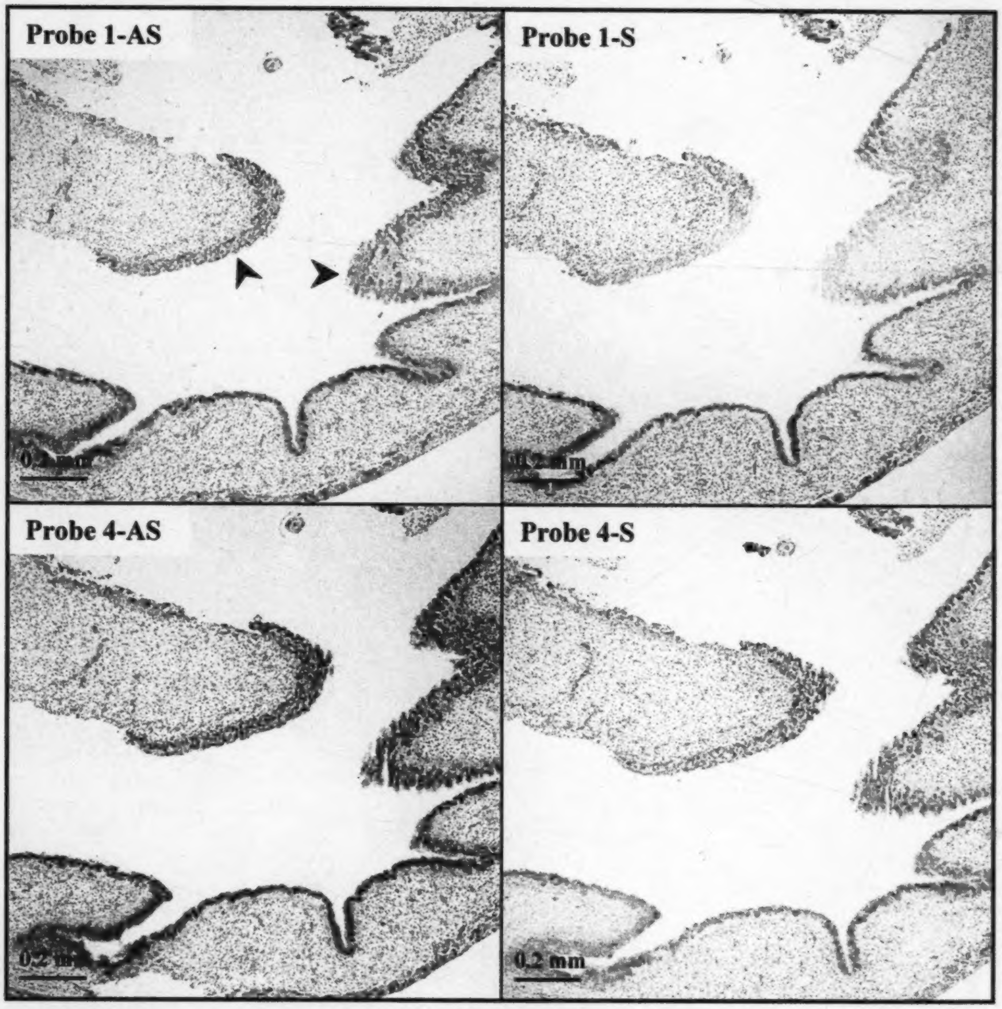
Probe 1-AS
Probe 1-S
Probe 4-AS
Probe 4-S
0.2 mm
0.2 mm
0.2 mm
0.2 mm

Figure 3-22. RNA *in situ* hybridization of human fetal week 8.5 sections of the stomach. Staining due to the various probes is indicated as a purple colour, while the counter-stain is green. Note the staining of the epithelial lining (arrowheads) when the 1-AS, 4-AS, and 4-S probes are used, while the section is negative with the 1-S probe.

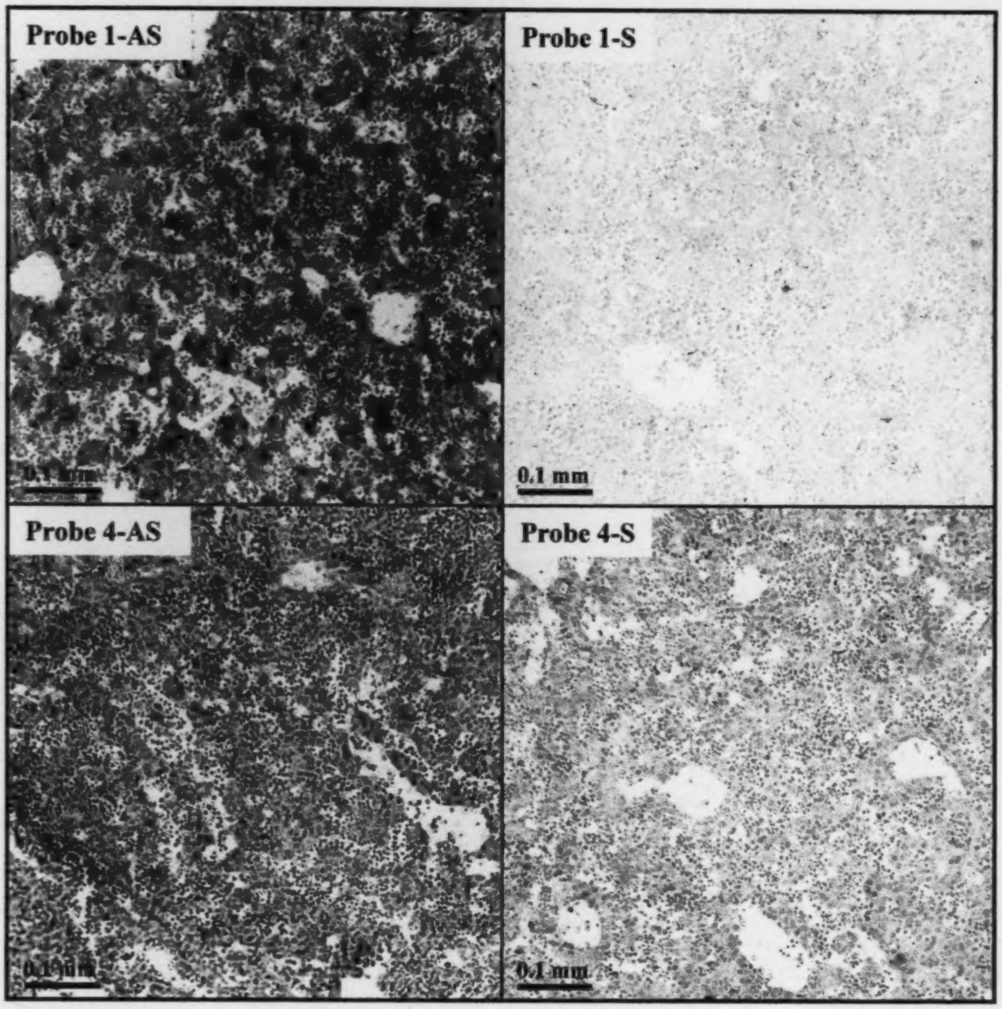
Probe 1-AS
Probe 1-S
Probe 4-AS
Probe 4-S
0.1 mm
0.1 mm
0.1 mm
0.1 mm

Figure 3-23. RNA *in situ* hybridization of human fetal week 10.7 liver sections. Staining due to the probe is indicated as a purple colour, while the counter-stain is green.

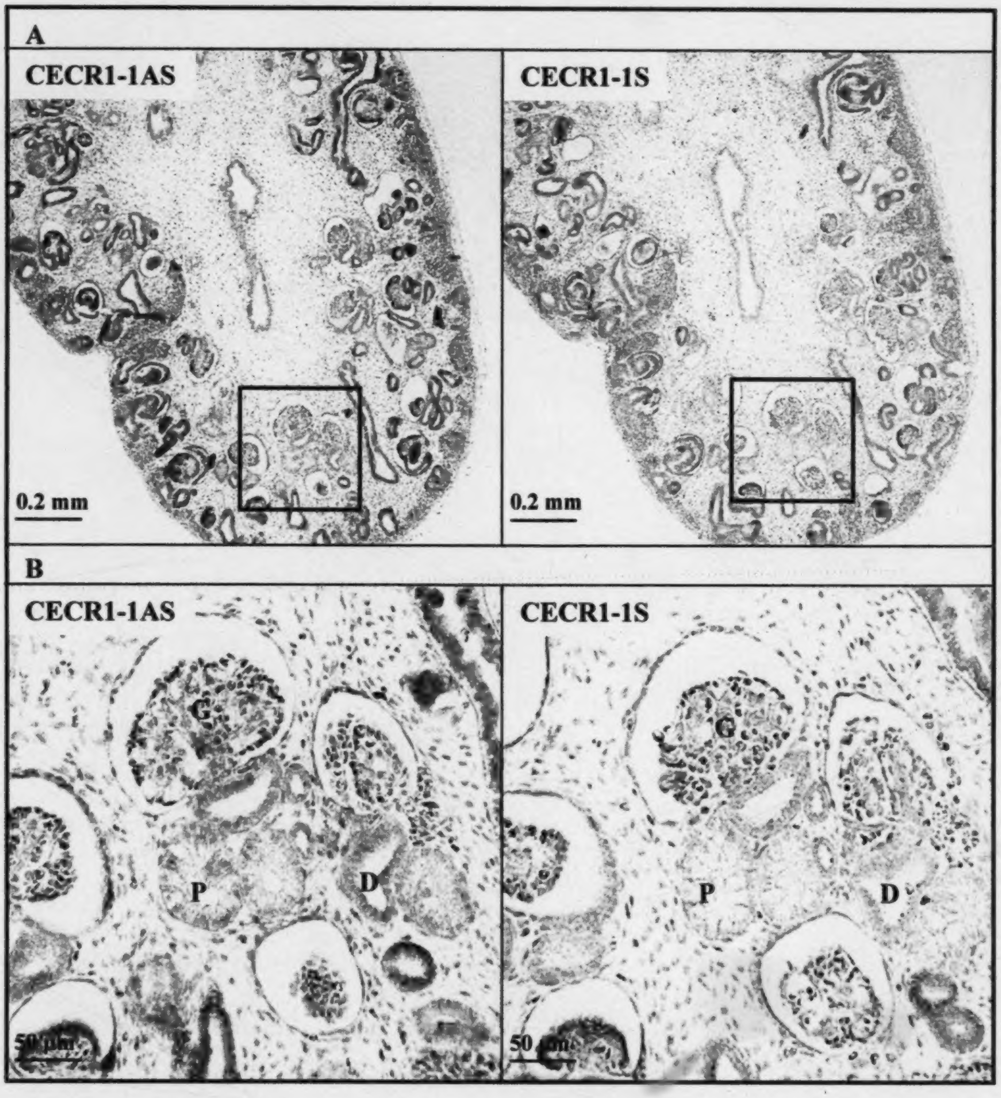
A
CECR1-1AS
CECR1-1S
0.2 mm
0.2 mm
B
CECR1-1AS
CECR1-1S
G
G
P
D
P
D
50 µm
50 µm

Figure 3-24. RNA *in situ* hybridization of human fetal week 10.7 metanephric kidney sections. **A)** Global view. **B)** Close-up view of the box shown in A. G, glomerulus; P, proximal convoluted tubule; D, distal convoluted tubule. Staining due to the probe is indicated as a purple/pink colour, while the counter-stain is green/blue.

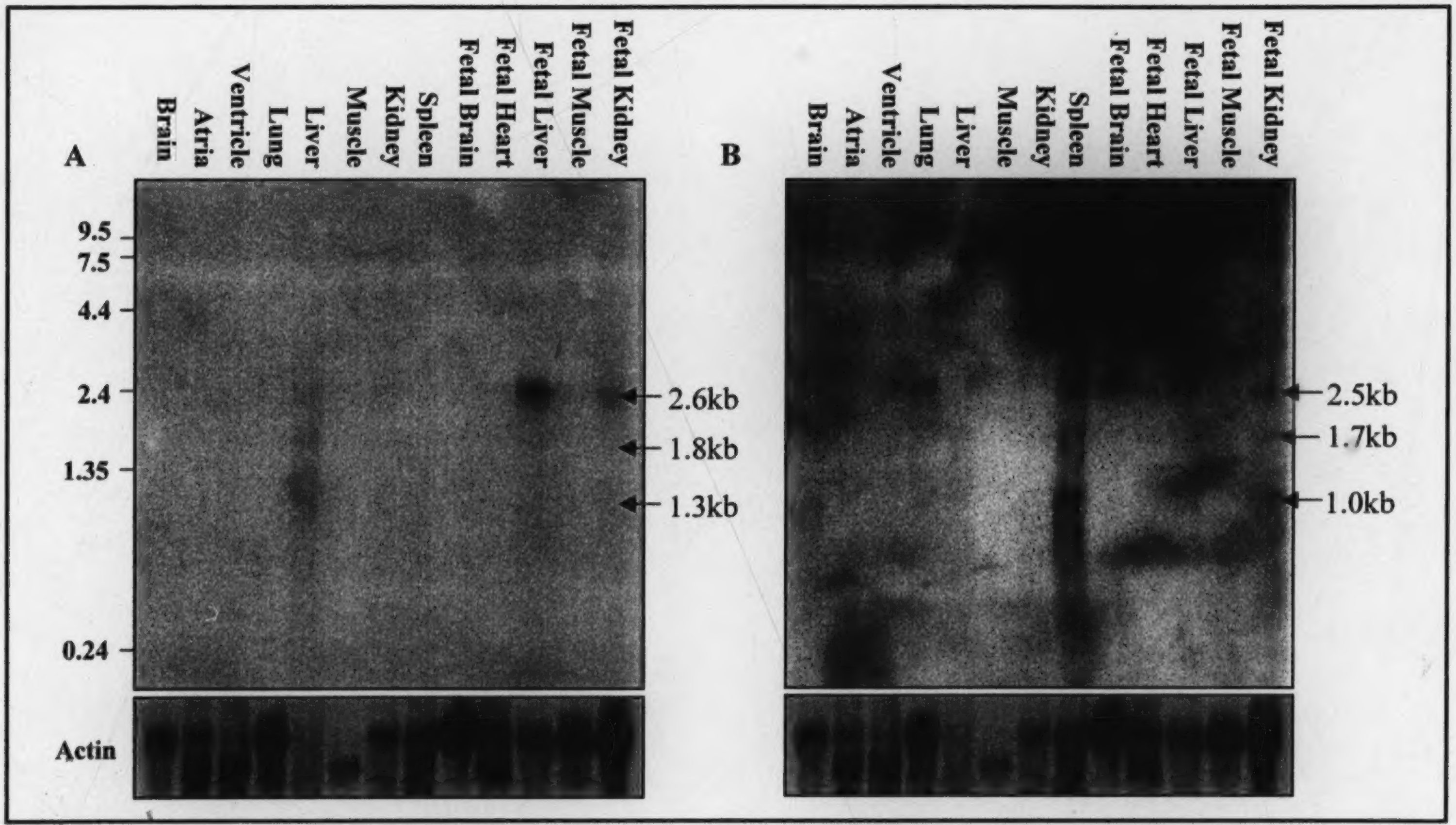
A
Brain
Atria
Ventricle
Lung
Liver
Muscle
Kidney
Spleen
Fetal Brain
Fetal Heart
Fetal Liver
Fetal Muscle
Fetal Kidney
9.5
7.5
4.4
2.4
1.35
0.24
Actin
2.6kb
1.8kb
1.3kb
B
Brain
Atria
Ventricle
Lung
Liver
Muscle
Kidney
Spleen
Fetal Brain
Fetal Heart
Fetal Liver
Fetal Muscle
Fetal Kidney
2.5kb
1.7kb
1.0kb

Figure 3-25. Expression profile of pig *CECR1* transcripts. A Northern blot was prepared from adult and fetal tissues, and probed with **A)** the CECR1-2S ssDNA probe to detect the antisense transcript, then **B)** a ssDNA probe from Probe A (Figure 3-11A) to detect the sense transcript . Multiple bands are detected in different tissues for each probe. The actin loading control is beneath.

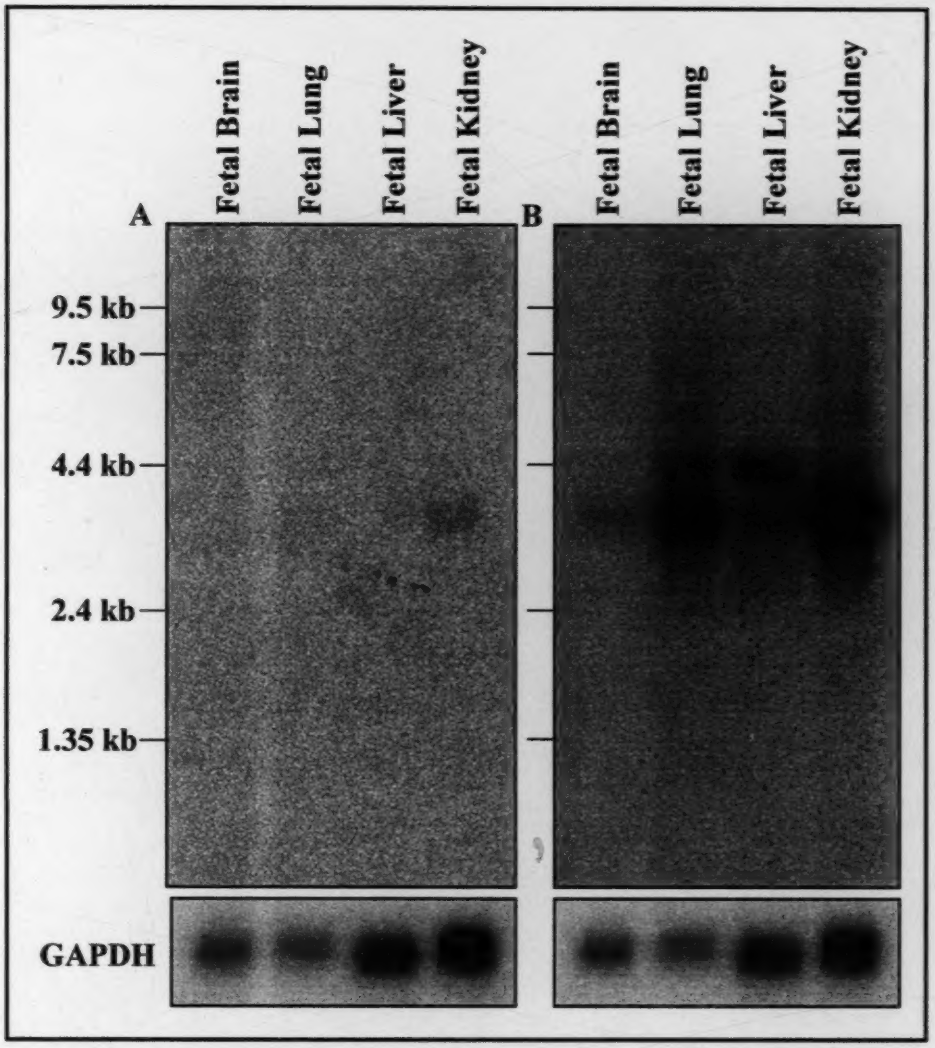
A
B
Fetal Brain
Fetal Lung
Fetal Liver
Fetal Kidney
Fetal Brain
Fetal Lung
Fetal Liver
Fetal Kidney
9.5 kb
7.5 kb
4.4 kb
2.4 kb
1.35 kb
GAPDH

Figure 3-26. Expression profile of the human *CECR1* antisense transcript. A Clontech fetal Northern blot (lot#7100219) was probed sequentially with **A)** the F7S ssDNA probe to detect the antisense transcript, or **B)** the R1AS ssDNA probe to detect the sense transcript. An antisense transcript of 3.4 kb is detected in the fetal kidney lane of blot A, whereas bands of 4.4 and 3.5 kb are detected in blot B. The stages represented in each tissue are as follows: brain (20-25wk, pool of 9), lung (18-28wk, pool of 29), liver (18-24wk, pool of 17), kidney (23-36wk, pool of 13). The GAPDH loading control is beneath each blot.

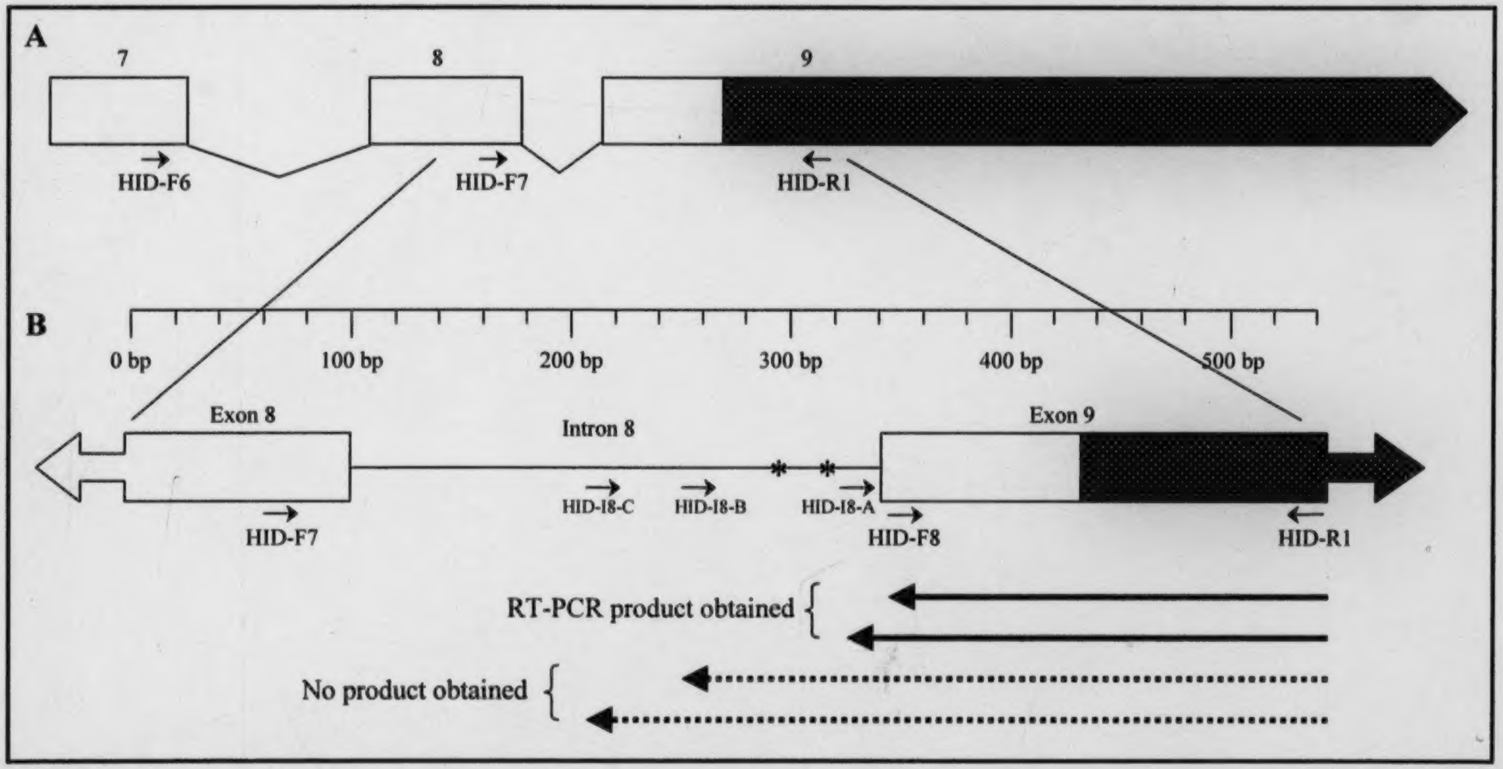
A
7
8
9
HID-F6
HID-F7
HID-R1
B
0 bp
100 bp
200 bp
300 bp
400 bp
500 bp
Exon 8
Intron 8
Exon 9
HID-F7
HID-I8-C
HID-I8-B
HID-I8-A
HID-F8
HID-R1
RT-PCR product obtained
No product obtained

Figure 3-27. Summary of results for the *CECR1* antisense transcript. **A)** The structure of *CECR1* exons 7-9, showing the primers used to make the RNA *in situ* hybridization and Northern hybridization probes from the *CECR1* cDNA. **B)** Close-up of intron 8 (depicted as a horizontal line), showing the "reverse" primers used to obtain the antisense RT-PCR products shown. Only the HID-F8 and HID-I8-A "reverse" primers gave rise to a product. Two potential splice donor sites (*) for the antisense transcript are also indicated.

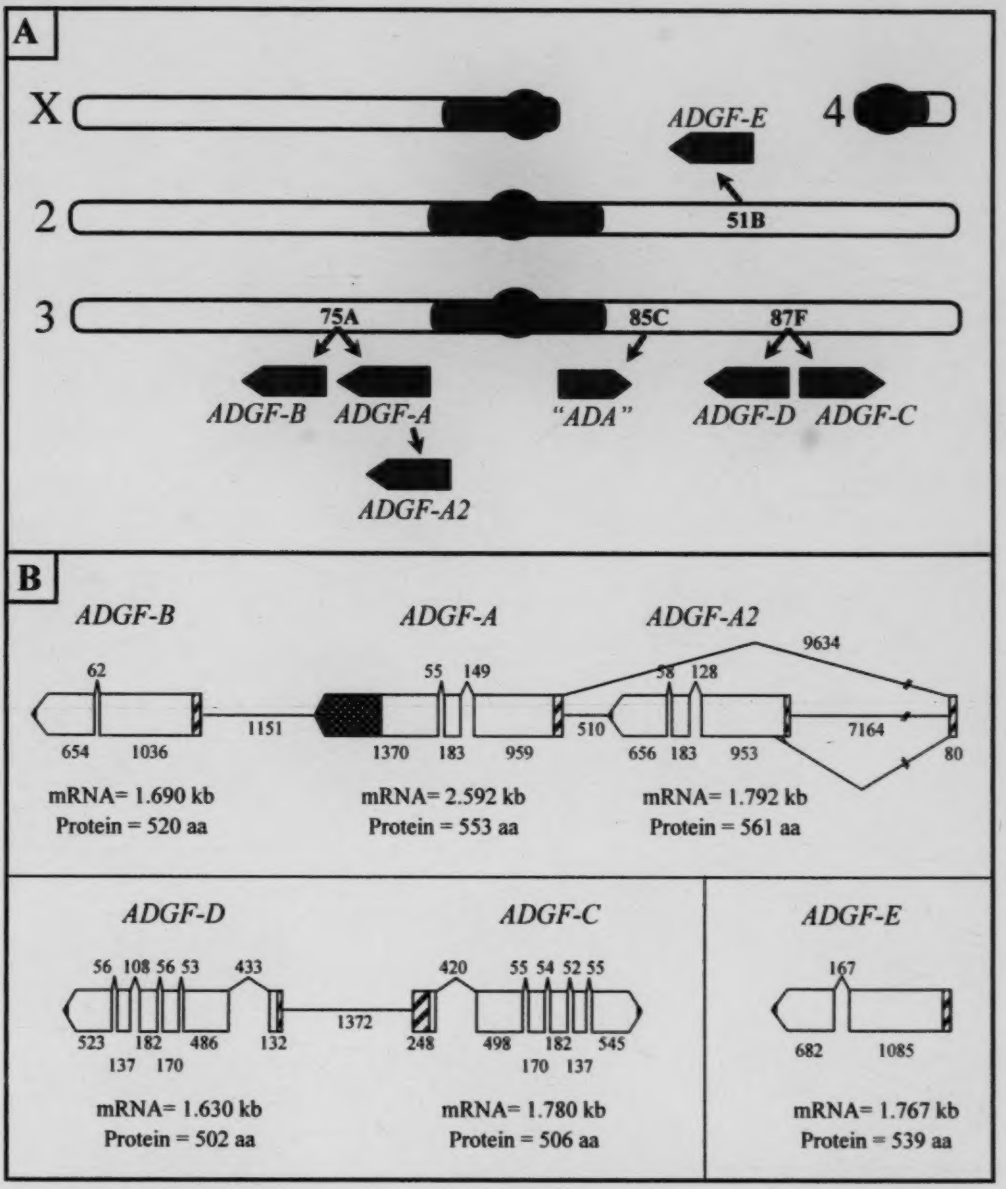
A
X
ADGF-E
4
2
51B
3
75A
85C
87F
ADGF-B
ADGF-A
"ADA"
ADGF-D
ADGF-C
ADGF-A2
B
ADGF-B
ADGF-A
ADGF-A2
9634
62
55
149
58
128
1151
510
7164
654
1036
1370
183
959
656
183
953
80
mRNA= 1.690 kb
Protein = 520 aa
mRNA= 2.592 kb
Protein = 553 aa
mRNA= 1.792 kb
Protein = 561 aa
ADGF-D
ADGF-C
ADGF-E
56 108 56 53
433
420
55 54 52 55
167
1372
523
182
486
132
137
170
248
498
182
545
170
137
682
1085
mRNA= 1.630 kb
Protein = 502 aa
mRNA= 1.780 kb
Protein = 506 aa
mRNA= 1.767 kb
Protein = 539 aa

Figure 3-28. Drosophila *ADGF* homologues. **A.** Chromosomal location and arrangement of the six *ADGF* genes; not to scale. The location of the Drosophila gene thought to be ADA (shown later to actually be ADAL) is also depicted. **B.** Genomic structure and inter-relatedness of the six genes. Distances are to scale. Numbers refer to base pairs, unless otherwise stated. Symbols denoting coding exons, 5' UTR, etc. are as in Figure 3-1. Horizontal lines represent intergenic distances.

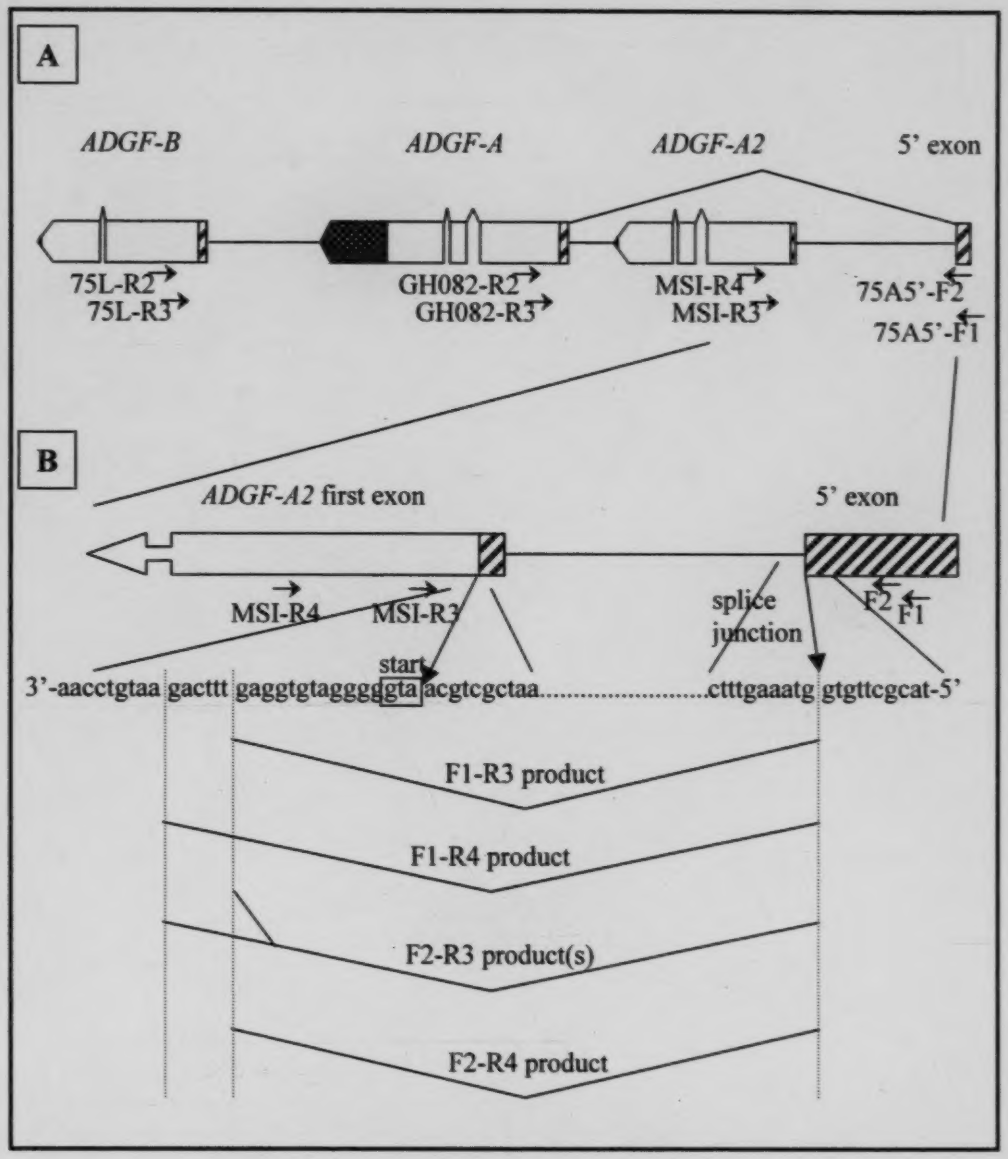
A
ADGF-B
ADGF-A
ADGF-A2
5' exon
75L-R2
75L-R3
GH082-R2
GH082-R3
MSI-R4
MSI-R3
75A5'-F2
75A5'-F1
B
ADGF-A2 first exon
5' exon
MSI-R4
MSI-R3
splice junction
F2
F1
start
3'-aacctgtaa gacttt gaggtgtagggggta acgtcgctaa........................ctttgaaatg gtgttcgcat-5'
F1-R3 product
F1-R4 product
F2-R3 product(s)
F2-R4 product

Figure 3-29. Summary of RT-PCR results in the 75A region. **A)** Locations of primers used in the study. Splicing of ADGF-A to the 5' exon was already established from the GH08276 cDNA. No products were found linking the 5' exon with ADGF-B. **B)** Focus on the four RT-PCR products obtained for ADGF-A2. Two different acceptor splice sites were utilized when splicing from the 5'exon. The published start codon (Matsushita et al., 2000) is boxed. The alternate start codon is 57 bp downstream and is not shown here.

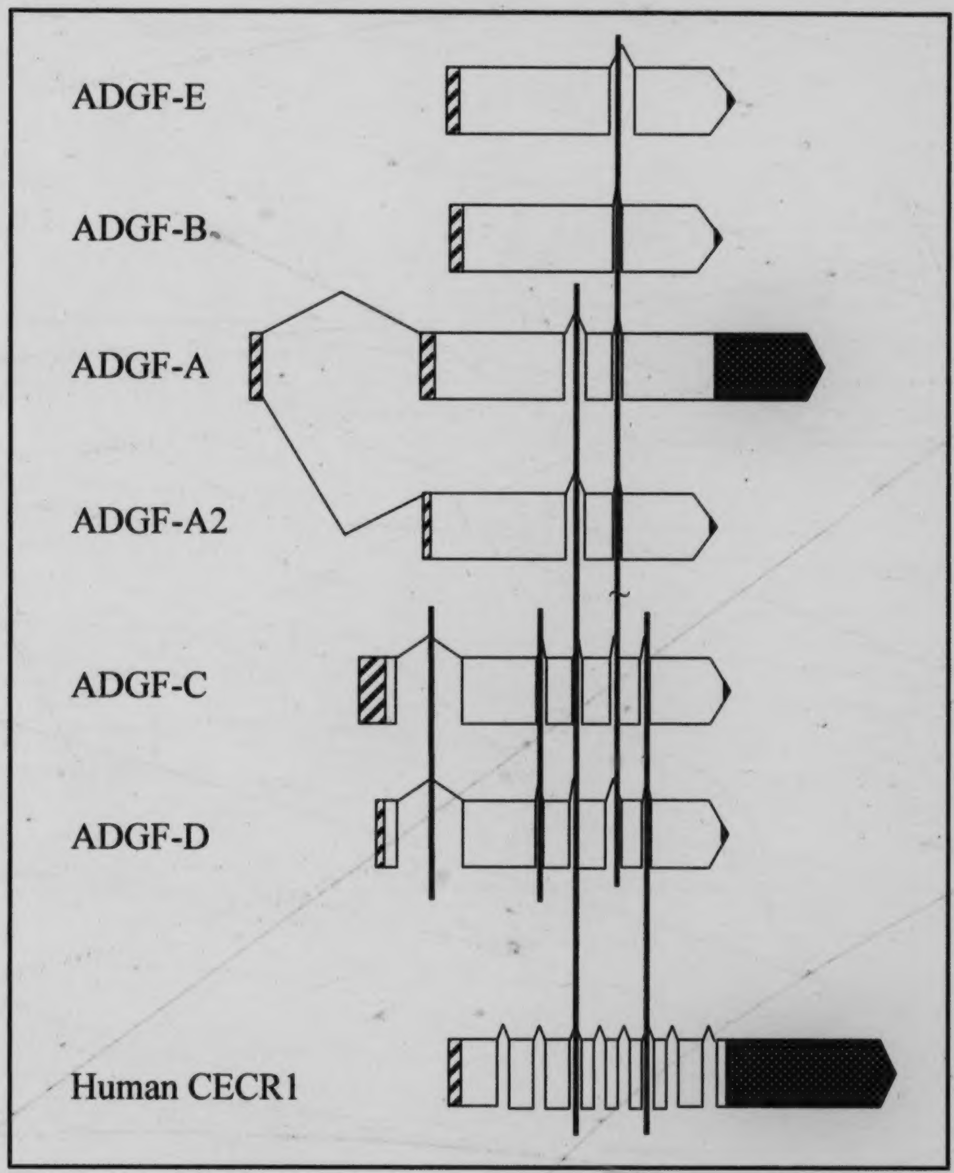
ADGF-E
ADGF-B
ADGF-A
ADGF-A2
~
ADGF-C
ADGF-D
Human CECR1

Figure 3-30. Schematic depiction of the intron alignment between the Drosophila *ADGF*s and human *CECR1*. Intron locations shared between genes are shown using vertical lines. The break (~) indicates a one base pair shift of the intron between two groups of genes. The human gene intron distances are not drawn to scale.

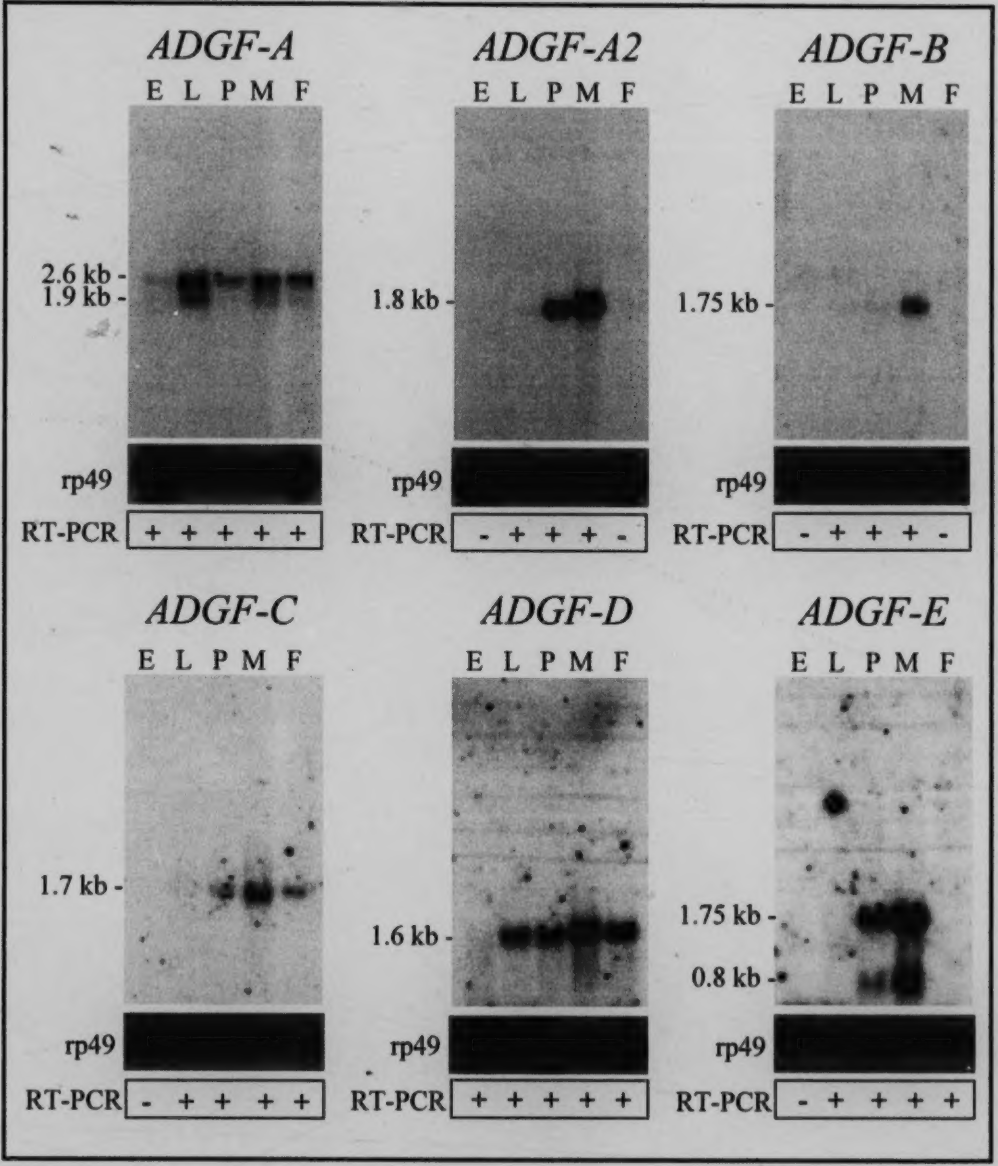
ADGF-A
E L P M F
2.6 kb -
1.9 kb -
rp49
RT-PCR + + + + +
ADGF-A2
E L P M F
1.8 kb -
rp49
RT-PCR - + + + -
ADGF-B
E L P M F
1.75 kb -
rp49
RT-PCR - + + + -
ADGF-C
E L P M F
1.7 kb -
rp49
RT-PCR - + + + +
ADGF-D
E L P M F
1.6 kb -
rp49
RT-PCR + + + + +
ADGF-E
E L P M F
1.75 kb -
0.8 kb -
rp49
RT-PCR - + + + +

Figure 3-31. Expression analysis of the Drosophila *ADGF* genes during development. Northern blots of Drosophila polyA (+) mRNA were made and probed by Lynn Podemski with the six *ADGF* genes. Lanes are embryo (E); larva (L); pupa (P); adult male (M); adult female (F). DNA probes were made from the following sources: (*ADGF-A*) PCR fragment from cDNA clone GH08276 using primers GH082-F & GH082-R (see Table 2-1); (*ADGF-A2*) PCR fragment from BAC clone 44L18 using

primers MSI-F & MSI-R; (*ADGF-B*) restriction fragment from BAC 44L18; (*ADGF-C*) restriction fragment from BAC 30G22; (*ADGF-D*) restriction fragment from BAC 30G22; (*ADGF-E*) restriction fragment from BAC 7P02. A restriction fragment of rp49 was used as a loading control for all blots. Detection of rare transcripts was carried out using RT-PCR. The primers for each gene were designed across introns and were as follows: (*ADGF-A*) primers GH082-F & GH082-R; (*ADGF-A2*) MSI-F2 & MSI-R2; (*ADGF-B*) 75A-L-F & 75A-L-R; (*ADGF-C*) LP055-F2 & LP055-R4; (*ADGF-D*) GH122-F & GH122-R; (*ADGF-E*) 50A-F & 50A-R (see Table 2-1). *RPS3a* served as a control for all stages, using primers RP5F-1 & RP3R-1. The (+) positive and (-) negative results are presented in table format under the Northern blots. PCR results were carried out in duplicate, and gave the same outcome.

Table 3-1. Names, accession numbers, and predicted signal peptides of proteins used in the study.

Protein Name[1]	Organism name	Common name	Gene accession[2]	Protein accession[2]	Supporting ESTs[3]	SP[4]
ADGF subfamily						
Hs_CECR1	*Homo sapiens*	human	AF190746	AAF65941		1-29
Pt_CECR1	*Pan troglodytes*	chimp	genomic AC135612 (A)		No ESTs	1-29
Pa_CECR1	*Papio anubis*	baboon	genomic AC091672 (A)		No ESTs	1-29
Ss_CECR1	*Sus scrofa*	pig	AF384216	AAL40921		1-24
Gg_CECR1	*Gallus gallus*	chicken	AY902779	AAX10953		1-23
Xl_CECR1	*Xenopus laevis*	frog	AY902778	AAX10952		1-19
Dr_CECR1-1	*Danio rerio*	zebrafish	AF384217	AAL40922		1-24
Dr_CECR1-2	*Danio rerio*	zebrafish	genomic BX323558 (A)		No ESTs	1-26
Tr_CECR1-1	*Takifugu rubripes*	pufferfish	Fugu scaffold_10227 (A)		No ESTs	1-25
Tr_CECR1-2	*Takifugu rubripes*	pufferfish	Fugu scaffold_1919 (A)		No ESTs	1-21
Tn_CECR1-1	*Tetraodon nigroviridis*	pufferfish	genomic CAAE01014566 (A)		No ESTs	1-25
Tn_CECR1-2	*Tetraodon nigroviridis*	pufferfish	genomic CAAE01014691 (A)		No ESTs	1-21
Ac_MDGF	*Aplysia californica*	sea slug	AF117336	AAD13112		1-25
Dm_ADGF-A	*Drosophila melanogaster*	fruit fly	AF337554	AAF49306		1-30
Dm_ADGF-A2	*Drosophila melanogaster*	fruit fly	AB025255	BAB18576		No
Dm_ADGF-B	*Drosophila melanogaster*	fruit fly	AF384215	AAF49307		No
Dm_ADGF-C	*Drosophila melanogaster*	fruit fly	AF337552	AAF54980		1-19
Dm_ADGF-D	*Drosophila melanogaster*	fruit fly	AF337553	AAF54979		1-22
Dm_ADGF-E	*Drosophila melanogaster*	fruit fly	AF337551	AAF58224		No
Dp_ADGF-A	*Drosophila pseudoobscura*	fruit fly	genomic AADE01002456 (A)		No ESTs	1-23
Dp_ADGF-A2	*Drosophila pseudoobscura*	fruit fly	genomic AADE01002456 (A)		No ESTs	No
Dp_ADGF-B	*Drosophila pseudoobscura*	fruit fly	genomic AADE01002456 (A)		No ESTs	No
Dp_ADGF-C	*Drosophila pseudoobscura*	fruit fly	genomic AADE01000100 (A)		No ESTs	1-19
Dp_ADGF-D	*Drosophila pseudoobscura*	fruit fly	genomic AADE01000100 (A)		No ESTs	1-19
Dp_ADGF-E	*Drosophila pseudoobscura*	fruit fly	genomic AADE01000620 (A)		No ESTs	No
Dy_ADGF-A	*Drosophila yakuba*	fruit fly	genomic AAEU01004459 (A)		No ESTs	1-30
Dy_ADGF-A2	*Drosophila yakuba*	fruit fly	genomic AAEU01004459 (A)		No ESTs	No
Dy_ADGF-B	*Drosophila yakuba*	fruit fly	genomic AAEU01004459 (A)		No ESTs	No
Dy_ADGF-C	*Drosophila yakuba*	fruit fly	genomic AAEU01000335 (A)		No ESTs	1-19
Dy_ADGF-D	*Drosophila yakuba*	fruit fly	genomic AAEU01000335 (A)		No ESTs	1-20
Dy_ADGF-E #	*Drosophila yakuba*	fruit fly	genomic AAEU01002956 (A)		No ESTs	N/A
Sp_ADGF-A	*Sarcophaga peregrina*	flesh fly	D83125	BAA11812		1-18
Gm_TSGF-1	*Glossina m. morsitans*	tsetse fly	AF140521	AAD52850		1-21
Gm_TSGF-2	*Glossina m. morsitans*	tsetse fly	AF140522	AAD52851		1-19
Ll_ADA	*Lutzomyia longipalpis*	sandfly	AF234182	AAF78901		1-18
Ag_ADGF-1	*Anopheles gambiae*	mosquito	XM_308848	XP_308848	BX623738	No
Ag_ADGF-2	*Anopheles gambiae*	mosquito	genomic AAAB01008810 (B)		No ESTs	1-20

Ag_ADGF-3	*Anopheles gambiae*	mosquito	genomic AAAB01008807 (B)		BX627955	No
Ag_ADGF-4 #	*Anopheles gambiae*	mosquito	genomic AAAB01002509 (A)		No ESTs	N/A
Cq_ADA	*Culex p. quinquefasciatus*	mosquito	AF298886	AAK97208		1-17
Aa_ADA	*Aedes aegypti*	mosquito	AF466610	AAL76033		1-26
Um_ADGF	*Ustilago maydis*	fungus	genomic AACP01000068 (B)		No ESTs	No
Nc_ADGF-1	*Neurospora crassa*	fungus	XM_323997	XP_323998	AW710270	No
Nc_ADGF-2	*Neurospora crassa*	fungus	XM_323366	XP_323367	BG279966	No
Gz_ADGF-1	*Gibberella zeae*	fungus	XM_390381	XP_390381	CD460809	No
Gz_ADGF-2	*Gibberella zeae*	fungus	XM_386598	XP_386598$	No ESTs	No
Mg_ADGF-1	*Magnaporthe grisea*	fungus	genomic AACU01001458 (C)		No ESTs	No
Mg_ADGF-2	*Magnaporthe grisea*	fungus	genomic AACU01001430 (B)		No ESTs	No
An_ADGF-1	*Aspergillus nidulans*	fungus	genomic AACD01000042 (C)		No ESTs	No
An_ADGF-2	*Aspergillus nidulans*	fungus	genomic AACD01000094 (B)		No ESTs	No
Dd_ADGF	*Dictyostelium discoideum*	mould	genomic AC116305 (C)		C89929	1-26
ADAL subfamily						
Hs_ADAL	*Homo sapiens*	human	XM_091156	XP_091156$	CR739704	No
Pt_ADAL	*Pan troglodytes*	chimp	genomic AADA01232690 (A)		No ESTs	No
Mm_ADAL	*Mus musculus*	mouse	BC052048	AAH52048		No
Rn_ADAL	*Rattus norvegicus*	rat	genomic NW_047657 (B)		CO393373	No
Ss_ADAL #	*Sus scrofa*	pig			BI343718	No
Gg_ADAL	*Gallus gallus*	chicken	genomic AADN01061886 (A)		AJ454771	No
Xl_ADAL	*Xenopus laevis*	frog	BC073685	AAH73685		No
Dr_ADAL #	*Danio rerio*	zebrafish			CN015078	No
Tr_ADAL	*Takifugu rubripes*	pufferfish	genomic CAAB01000380 (A)		No ESTs	No
Tn_ADAL	*Tetraodon nigroviridis*	pufferfish	genomic CAAE01015000 (B)		No ESTs	No
Dm_ADA	*Drosophila melanogaster*	fruit fly	NM_141609	NP_649866	BI213048	No
Dp_ADAL	*Drosophila pseudoobscura*	fruit fly	genomic AADE01000441 (A)		No ESTs	No
Dy_ADAL	*Drosophila yakuba*	fruit fly	genomic AAEU01001954 (A)		No ESTs	No
Ag_ADAL	*Anopheles gambiae*	mosquito	genomic AAAB01008900 (B)		No ESTs	No
Ce_ADAL	*Caenorhabditis elegans*	worm	NM_182155	NP_871955	BJ103876	No
Um_ADAL	*Ustilago maydis*	fungus	XM_398179	XP_398179	No ESTs	No
Nc_ADAL	*Neurospora crassa*	fungus	XM_322523	XP_322524$	No ESTs	No
Gz_ADAL #	*Gibberella zeae*	fungus	genomic AACM01000179 (B)		No ESTs	No
An_ADAL	*Aspergillus nidulans*	fungus	genomic AACD01000010 (B)		CK448224	No
ADA subfamily						
Hs_ADA	*Homo sapiens*	human	NM_000022	NP_000013	BC040226	No
Pt_ADA	*Pan troglodytes*	chimp	genomic AADA01316146 (A)		No ESTs	No
Mm_ADA	*Mus musculus*	mouse	BC002075	AAH02075		No
Rn_ADA	*Rattus norvegicus*	rat	AB059655	BAB69691		No
Ss_ADA #	*Sus scrofa*	pig			BI337990	N/A
Gg_ADA	*Gallus gallus*	chicken	genomic AADN01030130 (A)		BU122720	No

Xl_ADA	*Xenopus laevis*	frog	BC073271	AAH73271		No
Dr_ADA	*Danio rerio*	zebrafish	BC076532	AAH76532		No
Tr_ADA #	*Takifugu rubripes*	pufferfish	genomic CAAB01001456 (A)		BU806270	No
Tn_ADA	*Tetraodon nigroviridis*	pufferfish	genomic CAAE01014729 (B)		No ESTs	No
Ce_ADA	*Caenorhabditis elegans*	worm	NM_182291	NP_872091	BJ771252	No
Ec_ADA	*Escherichia coli*	bacteria	M59033	AAA23419		No
Sco_ADA	*Streptomyces coelicolor*	bacteria	NC_003888	CAC33066	No ESTs	No
ADE subfamily						
Sce_ADE	*Saccharomyces cerevisiae*	yeast	NC_001146	NP_014258		No
Gz_ADE	*Gibberella zeae*	fungus	XM_381743	XP_381743		No
An_ADE	*Aspergillus nidulans*	fungus	AF123460	AAL56636		No
Sco_ADE	*Streptomyces coelicolor*	bacteria	NC_003888	CAB66224		No
AMPD subfamily						
Hs_AMPD1	*Homo sapiens*	human	NM_000036	NP_000027		No
Hs_AMPD2	*Homo sapiens*	human	M91029	AAA62127		No
Hs_AMPD3	*Homo sapiens*	human	NM_000480	NP_000471		No
Mm_AMPD2	*Mus musculus*	mouse	AK004759	BAB23540		No
Mm_AMPD3	*Mus musculus*	mouse	BC040366	AAH40366		No
Rn_AMPD1	*Rattus norvegicus*	rat	NM_138876	NP_620231		No
Rn_AMPD3	*Rattus norvegicus*	rat	NM_031544	NP_113732		No
Gg_AMPD3	*Gallus gallus*	chicken	XM_420973	XP_420973		No
Dr_AMPD1	*Danio rerio*	zebrafish	BC063996	AAH63996		No
Dr_AMPD3	*Danio rerio*	zebrafish	NM_199848	NP_956142		No
Dm_AMPD	*Drosophila melanogaster*	fruit fly	NM_167385	NP_727740		No
Ag_AMPD	*Anopheles gambiae*	mosquito	XM_310496	XP_310496$		No
Ce_AMPD	*Caenorhabditis elegans*	worm	NM_062573	NP_494974		No
An_AMPD	*Aspergillus nidulans*	fungus	XM_413009	XP_413009		No
Dd_AMPD	*Dictyostelium discoideum*	mould	AF238311	AAF65407		No

[1] Genes were categorized into the ADGF, ADAL, ADA, ADE, or AMPD subfamilies based on protein sequence similarity to the associated human member. # - the full protein sequence could not be determined and the protein was therefore not used in the phylogenetic analyses.

[2] Accession numbers that include an underscore represent sequences that have been predicted and assembled by a database curator. $ - The protein sequence was altered to be used in the phylogenetic analysis. The word "genomic" preceding an accession number indicates that the genomic sequence represented by the accession number was used, either by the author (A) or by a database curator (C) or a combination of both (B, meaning that the prediction by the curator was altered by the author), to predict the associated protein sequence.

[3] Predicted genes whose expression is supported partially by the existence of at least one EST have that ESTs accession number listed, otherwise "No ESTs" is listed, indicating no expression support.

[4] The presence of a predicted signal peptide (SP) is indicated by the amino acid residues suspected to be cleaved off. "No" indicates that a signal sequence was not predicted. N/A - not applicable (a signal peptide could not be predicted because no start codon was found).

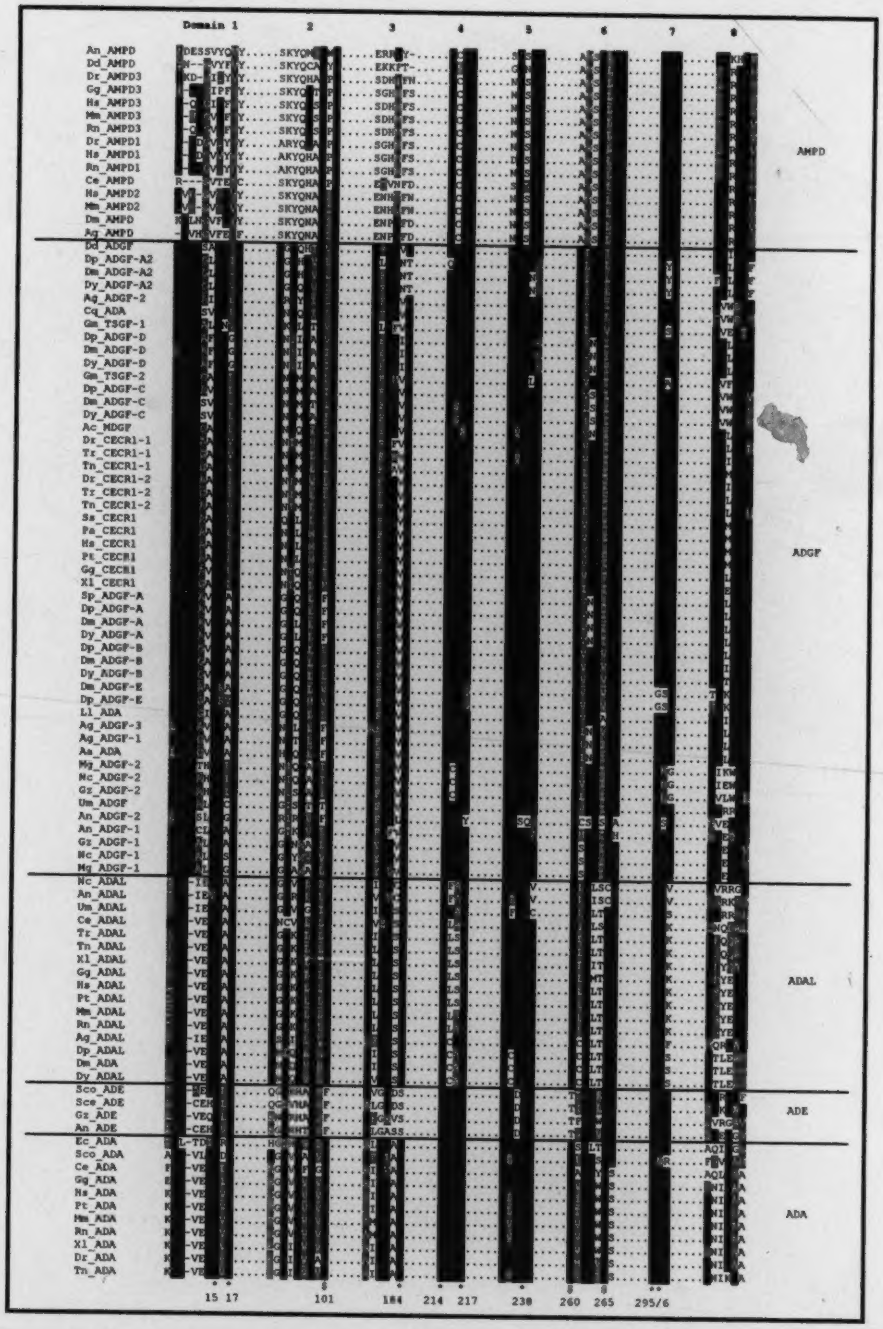
Domain 1
2
3
4
5
6
7
8
AMPD
ADGF
ADAL
ADE
ADA
15 17
101
184
214 217
238
260 265
295/6

Figure 3-32. Alignment and conserved domains in the adenyl-deaminase family. Background indicates identity (black) or conservative substitutions (gray). The amino acids important for ADA activity (*) or salt bridges ($) are numbered below according to mouse ADA. Horizontal lines delineate boundaries between subgroups. Species abbreviations are as noted in Table 3-1.

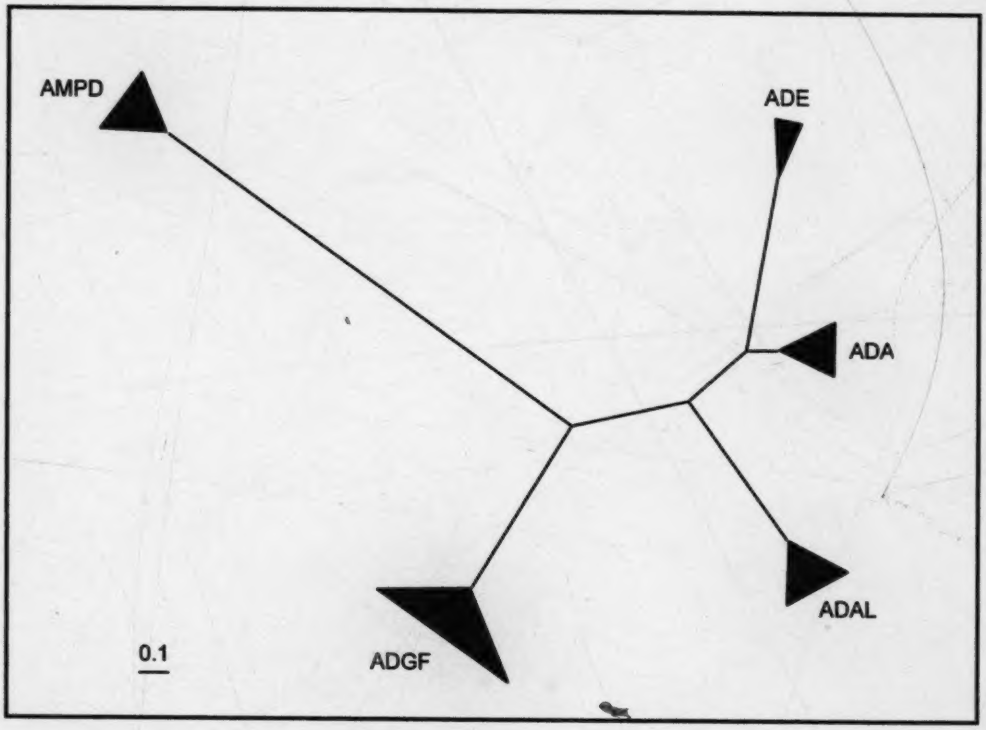

Figure 3-33. Initial analysis of the five protein subfamilies using MrBayes. This is a simplified representation of the tree inferred using all 95 of the ADGF, ADAL, ADA, ADE and AMPD proteins, showing the AMPD group as a natural outgroup. All taxa from each group were clipped from the tree and replaced by a triangle, whose width is proportional to the number of taxa in that group. The scale bar represents 0.1 substitutions per site.

Table 3-2. Summary of temperature settings and acceptance values for individual initial (150,000 generations) and final (550,000 generations) MrBayes analyses of the ingroup alignment. Trials containing temperature acceptance values outside of the normal range (10-70%) are not reliable.

Trial number	Temperature setting	Proportion of successful state exchanges between chains separated by one heating step[1]		
150,000 generations		Chains 1 and 2	Chains 2 and 3	Chains 3 and 4
1	0.2 (default)	16%	23%	29%
2	0.2	15%	23%	0%
3	0.1	29%	1%	0%
4	0.05	59%	1%	0%
5	0.01	57%	3%	3%
6	0.001	81%	58%	40%
7	0.5	1%	3%	3%
8	0.005	73%	9%	2%
9A	0.2	16%	0%	0%
9B	0.2	0%	0%	11%
9C	0.2	15%	20%	6%
550,000 generations				
10A	0.2	18%	24%	20%
10B	0.2	18%	7%	0%
10C	0.2	17%	23%	25%
10D	0.2	10%	11%	0%
10E	0.2	18%	25%	30%

[1] Numbers shaded in gray lie outside the range of 10% to70%.

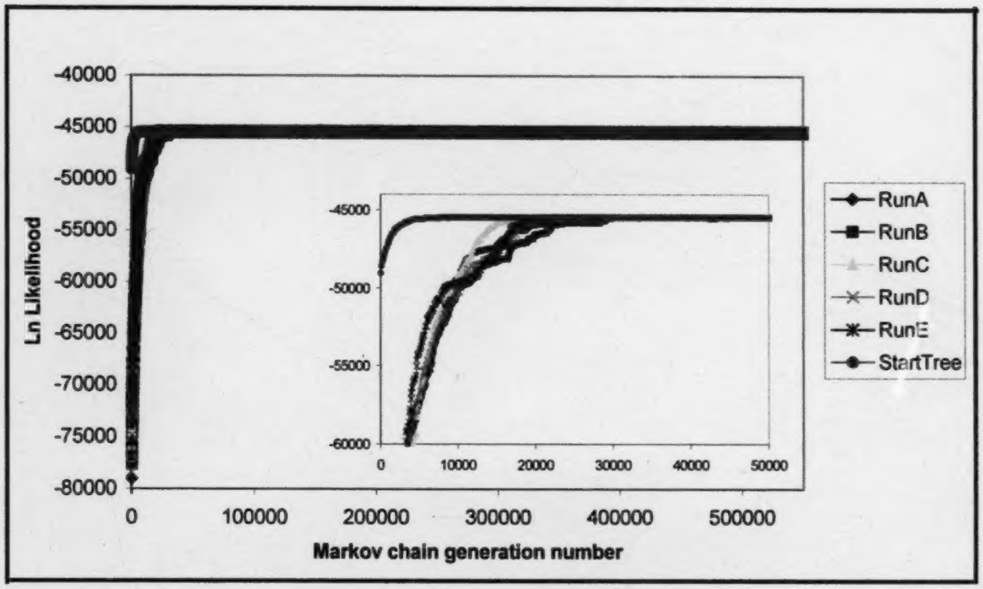

Figure 3-34. Convergence to stationarity during Bayesian phylogenetic analysis of the ingroup for individual MrBayes runs. Each of the five final runs (500,000 generations each, runs A-E) were plotted. The "StartTree" run used the maximum parsimony topology as the user defined starting tree (see text). Inset: Focus on the first 50,000 generations (burn-in) for the individual runs. Note that each run converged quickly to stationarity, meaning that the cold chain had reached a highly probable tree topology. Also, only the first 30,000 generations needed to be discarded as burn-in, since all runs had converged just prior to that point, but the first 50,000 generations were discarded to be safe.

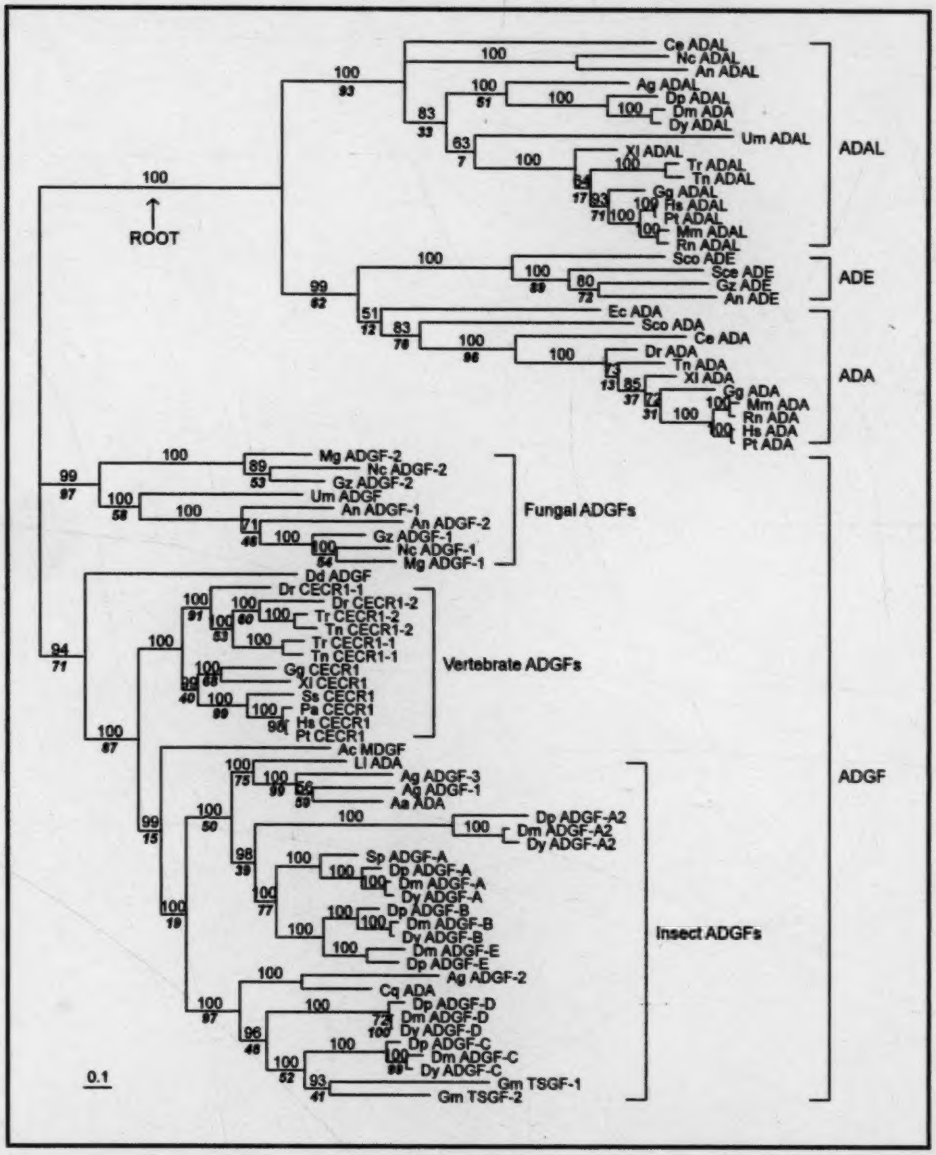
Ce ADAL
Nc ADAL
An ADAL
Ag ADAL
Dp ADAL
Dm ADA
Dy ADAL
Um ADAL
Xl ADAL
Tr ADAL
Tn ADAL
Gg ADAL
Hs ADAL
Pt ADAL
Mm ADAL
Rn ADAL
ADAL
ROOT
Sco ADE
Sce ADE
Gz ADE
An ADE
ADE
Ec ADA
Sco ADA
Ce ADA
Dr ADA
Tn ADA
Xl ADA
Gg ADA
Mm ADA
Rn ADA
Hs ADA
Pt ADA
ADA
Mg ADGF-2
Nc ADGF-2
Gz ADGF-2
Um ADGF
An ADGF-1
An ADGF-2
Gz ADGF-1
Nc ADGF-1
Mg ADGF-1
Fungal ADGFs
Dd ADGF
Dr CECR1-1
Dr CECR1-2
Tr CECR1-2
Tn CECR1-2
Tr CECR1-1
Tn CECR1-1
Gg CECR1
Xl CECR1
Ss CECR1
Pa CECR1
Hs CECR1
Pt CECR1
Vertebrate ADGFs
Ac MDGF
Ll ADA
Ag ADGF-3
Ag ADGF-1
Aa ADA
Dp ADGF-A2
Dm ADGF-A2
Dy ADGF-A2
Sp ADGF-A
Dp ADGF-A
Dm ADGF-A
Dy ADGF-A
Dp ADGF-B
Dm ADGF-B
Dy ADGF-B
Dm ADGF-E
Dp ADGF-E
Ag ADGF-2
Cq ADA
Dp ADGF-D
Dm ADGF-D
Dy ADGF-D
Dp ADGF-C
Dm ADGF-C
Dy ADGF-C
Gm TSGF-1
Gm TSGF-2
Insect ADGFs
ADGF
0.1

Figure 3-35. Phylogenetic analysis of the ingroup using MrBayes. This unrooted tree was inferred from the alignment of the ADGF, ADAL, ADA and ADE gene products. The arrow indicates the approximate location of the root if the AMPD outgroup had been included. The scale bar represents 0.1 substitutions per site. Species abbreviations are as noted in Table 3-1. Posterior probabilities are depicted as percentages on top of the internal branches, while MP bootstrap proportions are indicated below the branches in bold italics for comparison (Figure 3-36). Bootstrap values identical to the Bayesian probabilities were not included in the figure.

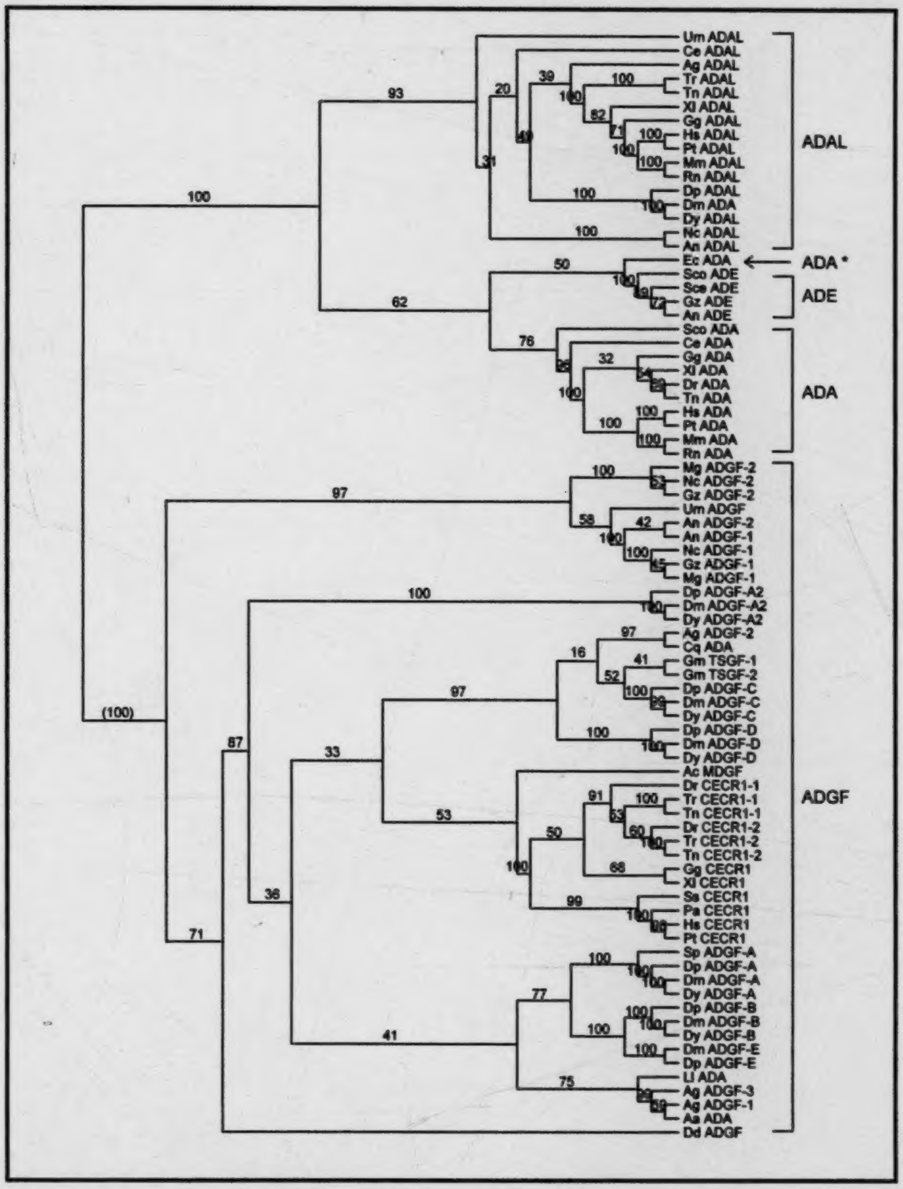
Um ADAL
Ce ADAL
Ag ADAL
Tr ADAL
Tn ADAL
Xl ADAL
Gg ADAL
Hs ADAL
Pt ADAL
Mm ADAL
Rn ADAL
Dp ADAL
Dm ADA
Dy ADAL
Nc ADAL
An ADAL
Ec ADA
Sco ADE
Sce ADE
Gz ADE
An ADE
Sco ADA
Ce ADA
Gg ADA
Xl ADA
Dr ADA
Tn ADA
Hs ADA
Pt ADA
Mm ADA
Rn ADA
Mg ADGF-2
Nc ADGF-2
Gz ADGF-2
Um ADGF
An ADGF-2
An ADGF-1
Nc ADGF-1
Gz ADGF-1
Mg ADGF-1
Dp ADGF-A2
Dm ADGF-A2
Dy ADGF-A2
Ag ADGF-2
Cq ADA
Gm TSGF-1
Gm TSGF-2
Dp ADGF-C
Dm ADGF-C
Dy ADGF-C
Dp ADGF-D
Dm ADGF-D
Dy ADGF-D
Ac MDGF
Dr CECR1-1
Tr CECR1-1
Tn CECR1-1
Dr CECR1-2
Tr CECR1-2
Tn CECR1-2
Gg CECR1
Xl CECR1
Ss CECR1
Pa CECR1
Hs CECR1
Pt CECR1
Sp ADGF-A
Dp ADGF-A
Dm ADGF-A
Dy ADGF-A
Dp ADGF-B
Dm ADGF-B
Dy ADGF-B
Dm ADGF-E
Dp ADGF-E
Ll ADA
Ag ADGF-3
Ag ADGF-1
Aa ADA
Dd ADGF
ADAL
ADA *
ADE
ADA
ADGF
93
100
62
97
(100)
100
87
97
33
53
36
41
71
39
20
31
82
71
76
32
58
42
16
41
52
91
50
68
99
77
75

Figure 3-36. Phylogenetic analysis of the ingroup using Maximum parsimony. This cladogram depicts the most parsimonious tree inferred using MP analysis of the ingroup alignment. Species abbreviations are as noted in Table 3-1. The bootstrap proportions from 100 replicates are shown on the internal branches. *E. coli* ADA (*) was placed with the ADE subgroup on this tree.

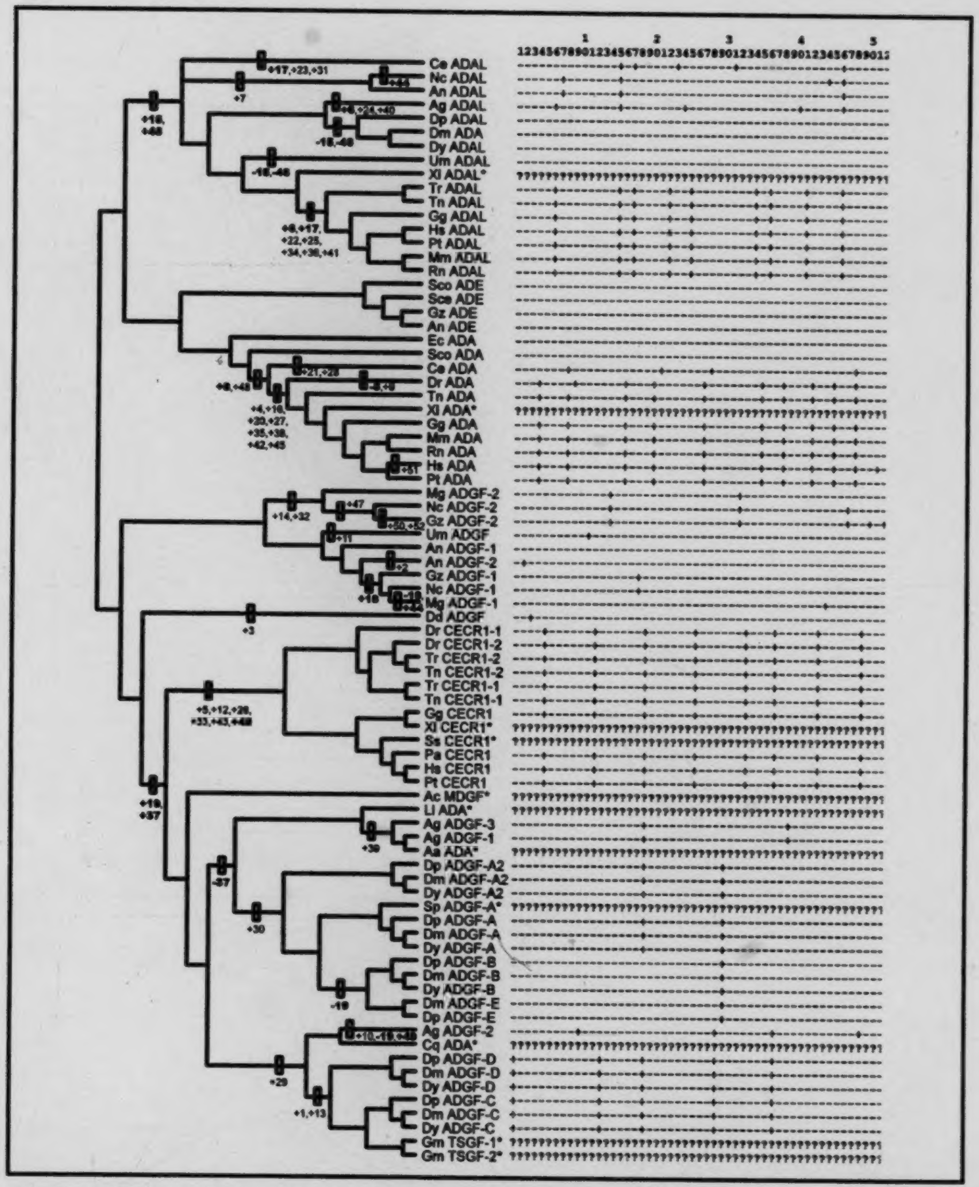
1 2 3 4 5
1234567890123456789012345678901234567890123456789012
Ce ADAL
Nc ADAL
An ADAL
Ag ADAL
Dp ADAL
Dm ADA
Dy ADAL
Um ADAL
Xl ADAL*
Tr ADAL
Tn ADAL
Gg ADAL
Hs ADAL
Pt ADAL
Mm ADAL
Rn ADAL
Sco ADE
Sce ADE
Gz ADE
An ADE
Ec ADA
Sco ADA
Ce ADA
Dr ADA
Tn ADA
Xl ADA*
Gg ADA
Mm ADA
Rn ADA
Hs ADA
Pt ADA
Mg ADGF-2
Nc ADGF-2
Gz ADGF-2
Um ADGF
An ADGF-1
An ADGF-2
Gz ADGF-1
Nc ADGF-1
Mg ADGF-1
Dd ADGF
Dr CECR1-1
Dr CECR1-2
Tr CECR1-2
Tn CECR1-2
Tr CECR1-1
Tn CECR1-1
Gg CECR1
Xl CECR1*
Ss CECR1*
Pa CECR1
Hs CECR1
Pt CECR1
Ac MDGF*
Ll ADA*
Ag ADGF-3
Ag ADGF-1
Aa ADA*
Dp ADGF-A2
Dm ADGF-A2
Dy ADGF-A2
Sp ADGF-A*
Dp ADGF-A
Dm ADGF-A
Dy ADGF-A
Dp ADGF-B
Dm ADGF-B
Dy ADGF-B
Dm ADGF-E
Dp ADGF-E
Ag ADGF-2
Cq ADA*
Dp ADGF-D
Dm ADGF-D
Dy ADGF-D
Dp ADGF-C
Dm ADGF-C
Dy ADGF-C
Gm TSGF-1*
Gm TSGF-2*
+17,+23,+31
+7
+16,
+48
-16,-48
-16,-48
+6,+17,
+22,+25,
+34,+36,+41
+21,+28
+8,+48
+4,+16,
+20,+27,
+35,+38,
+42,+45
+51
+47
+14,+32
+50,+52
+11
+2
+16
+3
+5,+12,+26,
+33,+43,+49
+19,
+37
+39
-37
+30
-19
+29
+1,+13

Figure 3-37. Evolution of introns within the ingroup. (*Left*) A cladogram of the Bayesian topology with numbered intron positions (within the alignment) that have been gained (+) or lost (-), mapped according to the most parsimonious reconstruction. Intron positions in **bold** involve more than one step, and are found more than once in the figure. The intron status is unknown for taxa (*) that lack genomic data. (*Right*) A table of occurrences of the 52 intron positions within the ingroup alignment, from which the reconstruction on the left was derived. The presence (+), absence (-), or unknown (?) status of each intron position is indicated beside each taxa.

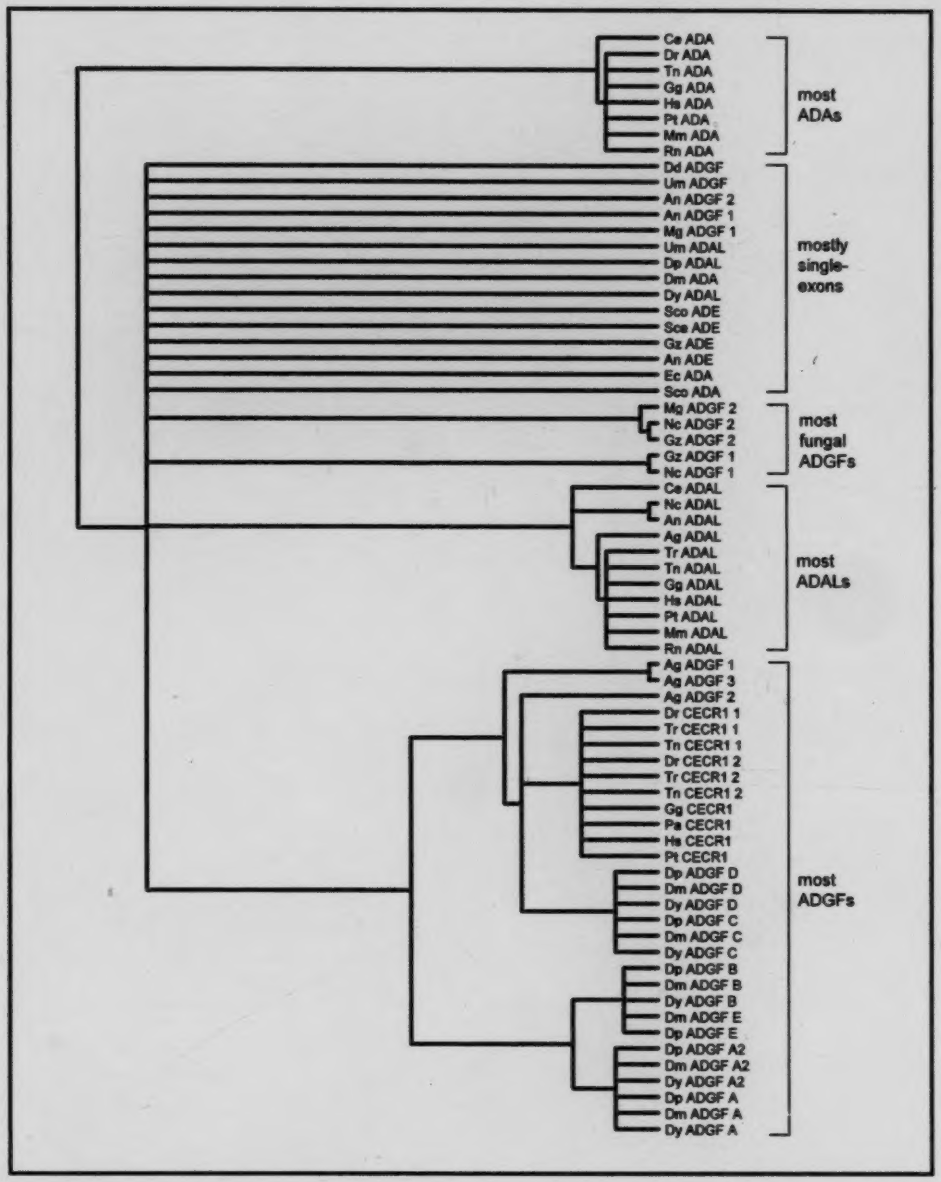
Ce ADA
Dr ADA
Tn ADA
Gg ADA
Hs ADA
Pt ADA
Mm ADA
Rn ADA
most ADAs
Dd ADGF
Um ADGF
An ADGF 2
An ADGF 1
Mg ADGF 1
Um ADAL
Dp ADAL
Dm ADA
Dy ADAL
Sco ADE
Sce ADE
Gz ADE
An ADE
Ec ADA
Sco ADA
mostly single-exons
Mg ADGF 2
Nc ADGF 2
Gz ADGF 2
Gz ADGF 1
Nc ADGF 1
most fungal ADGFs
Ce ADAL
Nc ADAL
An ADAL
Ag ADAL
Tr ADAL
Tn ADAL
Gg ADAL
Hs ADAL
Pt ADAL
Mm ADAL
Rn ADAL
most ADALs
Ag ADGF 1
Ag ADGF 3
Ag ADGF 2
Dr CECR1 1
Tr CECR1 1
Tn CECR1 1
Dr CECR1 2
Tr CECR1 2
Tn CECR1 2
Gg CECR1
Pa CECR1
Hs CECR1
Pt CECR1
Dp ADGF D
Dm ADGF D
Dy ADGF D
Dp ADGF C
Dm ADGF C
Dy ADGF C
Dp ADGF B
Dm ADGF B
Dy ADGF B
Dm ADGF E
Dp ADGF E
Dp ADGF A2
Dm ADGF A2
Dy ADGF A2
Dp ADGF A
Dm ADGF A
Dy ADGF A
most ADGFs

Figure 3-38. Maximum parsimony analysis of intron positions. A MP heuristic search was performed with the binary intron presence/absence matrix shown in Figure 3-37 (*right*), and the resulting topology is presented here. Note that the major groups are generally retained when considering only the conserved intron positions.

Chapter 4: Discussion

Human *IL-17* Receptor

The *IL-17R* gene resides at cytogenetic location 22q11.22-23 within the CES critical region and encodes a T-cell derived cytokine receptor with a broad tissue distribution (Yao et al., 1997). Two EST clusters were found distal to the coding sequence of *IL-17R*, which suggested possible alternative 3' ends for *IL-17R*. The use of both 3' ends in *IL-17R* transcripts was indeed confirmed by RT-PCR (Figure 3-1) and Northern blot analysis (Figure 3-2).

Since the completion of this work, it has been discovered that there are actually six IL-17 ligands, named IL-17A – IL-17F, each with a different associated expression pattern and receptor affinity (reviewed in Witowski et al., 2004). Also, besides IL-17R, four other IL-17 receptors have been found: IL-17RB – IL-17RE (reviewed in Kolls and Linden, 2004). The binding of IL-17A to IL-17R in particular has been characterized extensively as a mediator of host defense due to the stimulation of cytokines that specifically recruit neutrophils (Witowski et al., 2004). The levels of IL-17A seem to be important, since its overexpression may aggravate inflammatory reactions and actually contribute to tissue injury (Witowski et al., 2004).

CES patients show no overt immune system abnormalities, thus it was thought that *IL-17R* would play no role in CES if overexpressed (Footz et al., 2001). But the overexpression studies with IL-17A ligand suggest that perhaps the ligand-receptor relationship is dosage sensitive. Increased levels of IL-17R might have an effect on inflammation and may disrupt proper tissue development. Thus, although it may be unlikely, there is still a possibility that overexpression of *IL-17R* has a subtle effect in CES patients, and *IL-17R* cannot be discounted until further work rules it out completely as a candidate.

Interpretation of animal studies in *CECR1* overexpression

CECR1 is a very promising candidate for the features of CES, due to its expression pattern in fetal tissues and sequence similarity to insect growth factors (Riazi et al., 2000). The overexpression of *CECR1* homologues in mouse as a model system was carried out to determine if any phenotypic features of CES could be linked to CECR1. Transgenic mice were created using the human BAC 609c6 and observed for CES features. Almost all of the *CECR1* transgenic mouse founders and their progeny were phenotypically normal. The human *CECR1* transgene was expressed, as shown in Figure 3-6, but it could not be determined whether protein was produced, since a good antibody was not available. The need to severely decrease the BAC injection concentration, along with unusually low success in obtaining founders, suggested that there might have been a toxic effect of *CECR1* that caused the loss of most of the pregnancies. Two scenarios could be at work here. Perhaps there was some sort of mutation that prevented protein production in the transgene that lead to the four founders, whereas the intact protein caused the *in utero* death of all other litters. No gross or fine mutations in the transgene extracted from the transgenic mice were found to support this theory, however. Also, it is highly improbable that three separately mutated BACs were inserted into three mice that became founders from one litter. Another possibility might be that the DNA itself and not the content (i.e. the *CECR1* sequence) was the cause of toxicity in the various other attempts at higher BAC concentrations. The two litters that actually gave rise to the four founders may have overcome that toxicity somehow, although it is not apparent what factors might be involved. On the other hand, since there is no *CECR1* homologue in mouse, it is difficult to predict what effects the overexpression of this gene would cause in the mouse. Clearly, the explanation of the lack of phenotype in the transgenic mice will be better resolved when a good CECR1 antibody can be used to determine whether the mice are making the CECR1 protein.

Since there are *CECR1* homologues in zebrafish (*CECR1-1* and *CECR1-2*), this model organism could also be useful to study the function of *CECR1*. Overexpression studies done recently by Nyssa Ritzel and Fang Yang (unpublished results) using a zebrafish *CECR1-1* construct also showed no change in phenotype. This indicates that either the overexpression of *CECR1-1* has no effect on development on this organism, or that *CECR1-1* is actually required for development but these two experiments just did not disrupt it in an observable manner, or that no protein was produced from the construct. Again, the distinction between these possibilities would be better resolved with an antibody to determine if the CECR1 protein is actually produced from the overexpression construct. The overexpression of zebrafish *CECR1-2* will be carried out by Fang Yang to determine if there is any effect on development. Also, although not directly applicable to the CES duplication syndrome, the observation of zebrafish development that is deficient in *CECR1-1* (and/or *CECR1-2*) would provide information as to the developmental requirement of these genes. These experiments are in progress currently (Fang Yang), using antisense morpholinos (Summerton and Weller, 1997) to knockdown the expression of the two zebrafish *CECR1* genes, separately and together.

CECR1 antibodies only reliably detect the control samples

Antibodies to CECR1 could be used with the transgenic mice to determine if protein was being produced, as well as to study subcellular localization and protein-protein interactions. Five CECR1 antibodies were produced during this project. The 0A1 antibody was raised against the HIDExp1 recombinant protein, while antibodies 2F6 & 2F2, and 2F1 & 2F3, were raised against peptides CECR1-Pep1 and CECR1-Pep2, respectively (Figure 3-7). Each CECR1 antibody reliably recognized its purified control protein, and the 2F6 and 2F2 peptide antibodies also detected the HIDExp1 protein. Faint bands of 59 and/or 56 kDa were found in human heart and kidney protein extracts, depending on which of the five antibodies was used, and a strong 59 kDa band in spleen was observed with the 0A1 antibody. These bands coincidentally matched perfectly with the predicted sizes for CECR1 (Figure 3-3 and Table 4-1), and with the tissues expressing the *CECR1* mRNA (Figure 3-4). The 59 kDa band was also present in normal mouse, however, and both the human and mouse spleen bands persisted through the competition assay (Figure 3-9), suggesting that these bands were not specific. The immunoprecipitation assay showed that the identity of both bands detected by the 2F6 antibody was likely vimentin, an intracellular cytoskeletal protein (Clarke and Allan, 2002).

The fact that all five antibodies could detect their respective control protein, however, suggests that each is capable, in theory, of detecting the CECR1 protein. Perhaps the CECR1 protein is not expressed in the tissues tested, or more likely is not expressed in sufficient amounts to be detected on a Western blot. If CECR1 functions as a growth factor like other ADGF proteins, it would need to be strictly controlled, and might therefore be expected to exist in low amounts, making it difficult to detect. Also, if CECR1 is secreted, as predicted by the SignalP program (see Table 3-1), there might be a better chance to find it in concentrated blood serum or other extracellular fluids, such as peripheral blood leukocytes (Figure 3-4). Alternately, although the antibodies can bind the control protein when it is the only protein species in the sample, they may not have high enough avidity for CECR1 when copious other proteins are present in the sample. Two new CECR1 peptide antibodies are currently being evaluated by Fang Yang in the McDermid lab.

Striking similarities between CECR1 and ADA2

The suggestion that the CECR1 protein might be found in serum in the previous section brings to mind the fact that ADA2 is found mainly in serum. As mentioned in the introduction, ADA2 is one of three ADA isoforms found in certain cell types, and it is especially important in human plasma where it is responsible for the majority of ADA activity (Hirschhorn and Ratech, 1980). It is likely that ADA2 is a member of the adenyl-deaminase family. Since at least some ADGF members exhibit ADA activity, and the cytological location of ADA2 is yet to be found, there may be a connection between these two protein groups.

A recent paper described the purification of the chicken ADA2 protein from liver extracts, and showed that the 110 kDa dimer (active form) has a monomer weight of 55 kDa, and is N-glycosylated (Iwaki-Egawa et al., 2004). The mature (without the predicted 23 aa signal sequence) chicken CECR1 protein (ADGF member) is predicted to have a molecular weight of 55.7 kDa. Also, there are two N-glycosylation sites predicted with high confidence in the chicken CECR1 protein, at positions 172 and 295. The first 12 N-terminal amino acids (TPLWSLMQDLMM) of chicken ADA2 were determined by Edman degradation (Iwaki-Egawa et al., 2004). The first 12 N-terminal amino acids of chicken CECR1 (TPLWEDRDSLMQ) are surprisingly similar to, but not identical, to those of chicken ADA2. Note that the first and last four residues of the CECR1 N-terminus are identical to the first eight residues of the chicken ADA2 protein, suggesting that there may have been either a sequencing or a cloning error in one of the two sequences. The concordance of the chicken CECR1 cDNA sequence (accession # AY902779) and the genomic sequence, however, suggests that the protein sequencing may more likely be in error.

Other evidence for the comparison of chicken ADA2 to CECR1 comes from the analysis of other ADGF proteins. Human *CECR1* is expressed highly in peripheral blood leukocytes and in spleen (Figure 3-4), which also harbours many leukocytes; cells that might secrete CECR1 into the blood plasma. Also, *S. peregrina* ADGF-A was observed as a homodimer with two potential N-glycosylation sites, and its ADA activity was inhibited by 2'-deoxycoformycin (DCF) (Homma et al., 1996; Homma et al., 2001), which also inhibits ADA2 (Niedzwicki and Abernethy, 1991). Alternatively, ADA1 is inhibited by EHNA, while this compound does not affect ADA2. Two *D. melanogaster* homologues (ADGF-A, and -D) have been demonstrated to have ADA activity (Zurovec et al., 2002). Studies of *D. melanogaster* ADGF-A transfected into insect cells in the McDermid lab showed that its ADA activity is not inhibited by EHNA (Twila Yobb and Rezika Zurch, unpublished results). Altogether, these observations suggest that the ADGF subfamily may indeed be the identity of the elusive ADA2 protein.

ADA2 has been isolated from the liver of chicken and amphibian, and from humans and marsupials (Hirschhorn and Ratech, 1980) but has not been observed in livers of higher mammals including pig, cow, monkey (*Macaca rhesus*), and mouse, among others (Ma and Fisher, 1969) by the methods used. This observation might seem to dispute the hypothesis that CECR1 and ADA2 are the same protein, since CECR1 has been found in pig and cow (see Table 3-1). The existence of a *CECR1* antisense transcript in both human and pig tissues (especially pig liver; Figure 3-14) suggests that if down-regulation of the sense transcript is occurring, there may not be any CECR1 protein being produced in the liver at certain time points. It would be interesting to determine whether or not a *CECR1* antisense transcript exists in chicken liver, since its absence would lend further support to this theory, because much higher levels of ADA2 are found in chicken liver compared with human (Iwaki-Egawa et al., 2004). Also, the opossum (*Didelphis virginiana*) shows both ADA1 and ADA2 activity, whereas the plasma and other tissues of mice and rats only contains ADA1 but not ADA2 activity (Niedzwicki et al., 1995), providing further evidence that perhaps the rodents have lost ADA2.

It has been proposed that ADA1 and ADA2 take part in a homeostatic mechanism that controls the levels of 2-deoxyadenosine as a "weapon" in monocytes/macrophages against offending intracellular microorganisms (Gakis et al., 1998). If CECR1 is indeed the identity of ADA2, then perhaps CECR1 is involved in the immune system. Since the immune system is not overtly affected in CES patients, this would suggest that CECR1 is not involved in producing any of the CES features when overexpressed. Alternately, since CECR1 is predicted to be secreted, it might function in the blood serum to regulate adenosine levels elsewhere in the body, either autonomously or by joining with some unknown binding partner.

During the preparation of this thesis, a 57 kDa protein showing ADA activity in the presence of EHNA was isolated and purified from human plasma (Zavialov and Engstrom, 2005). Peptide sequencing of this human ADA2 protein revealed that it matched the protein sequence of human CECR1 (Zavialov and Engstrom, 2005), confirming that ADA2 and CECR1 are indeed the same protein as was hypothesized above. Since ADA2 activity is only a minor component of total ADA activity in human cells, and is much lower than chicken liver levels (Hirschhorn and Ratech, 1980; Iwaki-Egawa et al., 2004), this could explain why the CECR1 protein was difficult to detect by Western analysis. Also, since ADA2/CECR1 has a lower affinity for adenosine and a pH optimum at lower than physiological levels (Hirschhorn and Ratech, 1980; Andreasyan et al., 2005), this could explain why it has been difficult to show that CECR1 has ADA activity in some organisms, including humans (Rezika Zurch and Twila Yobb, unpublished results). A much higher adenosine concentration than required for ADA was used to show that human CECR1 has ADA activity (Andrey Zavialov, personal communication).

RNA *in situ* hybridization experiments are consistent with CES features

Since the expression pattern of the CECR1 protein could not be tested, the expression of the mRNA was examined. The RNA *in situ* hybridization results for the pig and human embryos were consistent both between species, and with many of the major tissues affected in CES patients. *CECR1* was expressed in a faint general pattern throughout the embryo in both species, with significant staining in the heart, kidney tubules, gut epithelium and liver. *CECR1* expression in heart, kidney and liver was also noted on the human fetal Northern blot (Figure 3-4), although the Northern blot represents older tissue (18-36 weeks, see Figure 3-26) compared with the sections. The Northern blot also showed strong expression of *CECR1* in fetal lung that was confirmed in the week 10.7 human embryo sections (Figure 3-21). The intensity of staining seemed to generally decrease in older embryo sections in both species. This trend is somewhat corroborated with the Northern data, in that both liver and kidney expression are dramatically reduced in the adult (Figure 3-4).

CECR1 seemed to be globally expressed in the embryonic heart. The two main heart defects in CES, total anomalous pulmonary venous return (TAPVR) and Tetralogy of Fallot (TOF), might be expected to result from a defect in signaling from the left atrium and pulmonary trunk, respectively. If *CECR1* is involved in the production of heart defects when overexpressed, it might be expected to be expressed only in specific locations, such as the left atrium or pulmonary artery wall, not the generalized staining pattern observed. This speculation, however, does not exclude *CECR1* as being involved in the production of heart defects in CES patients, since the localization of the CECR1 protein must still be elucidated.

CECR1 expression in the human day 34 heart as revealed by *in situ* hybridization (Figure 3-17) seemed to be stronger with the CECR1-1AS probe than the -4AS probe. Although this might be explained by differences in probe strength and hybridization time, since the sections were stained in two different experiments, this pattern actually contradicts the expression pattern of *CECR1* variant 1 versus variant 2. *CECR1v2* is expressed highly in the adult and fetal heart, while the variant 1 transcript is barely detectable in these tissues on the Northern blot (Figure 3-4). This might also be due to differences in ages on the Northern blot and embryo sections, but since probe 1AS only recognizes variant 1, while probe 4AS recognizes both variants, it might be expected that the signal from the 4AS probe should be stronger. Therefore the stronger staining intensity with the CECR1-1AS probe is most probably just a chance occurrence.

Expression of *CECR1* was also detected in blood cells within the heart atrium with the CECR1-1AS probe on the human fetal day 34 section (Figure 3-17). It was suggested in the results section that it was unclear as to whether this signal was artifactual. Since CECR1 is the same protein as ADA2, however, it would be expected to be expressed from blood cells, specifically monocytes/macrophages, such that the *in situ* result might not actually be an artifact.

CES patients often have kidney malformations, usually an absent/hypoplastic or polycystic kidney (Rosias et al., 2001). Human RNA *in situ* hybridization of week 10.7 sections showed that *CECR1* was expressed in the proximal convoluted tubules of the kidney, and the cells lining these tubules are known to excrete substances into the lumen (Cormack, 1993). If the CECR1 protein is secreted into the filtrate of the kidney, it may be able to act on other cells along the nephron or the entire length of the excretion pathway, which may be important for kidney development and/or function. The liver was also stained darkly with the *CECR1* antisense probes, as observed in both the day 34 (Figure 3-16 and 3-17) and day 47 (Figure 3-19) sections, although CES patients do not often suffer liver problems, besides very rare occurrences of liver failure due to biliary atresia (Rosias et al., 2001). Nonetheless, *CECR1* may be important for normal liver development or function.

Expression of *CECR1* was observed in the epithelial lining of the gut in pig embryos at day 20 and 28 (Figure 3-13 and 3-14). *CECR1* expression was also observed in the epithelial lining of the small bowel (day 34; Figure 3-17), pancreas (day 47; Figure 3-19), and stomach (week 8.5; Figure 3-22) of human embryo sections. This expression pattern may be related to CES, since patients show occasional instances of gut malrotation (Rosias et al., 2001), which often results from problems in signaling between the endodermally derived epithelium and the surrounding mesoderm (Ramalho-Santos et al., 2000). In normal development, *CECR1* may be important for proper gut development and/or is secreted into the gut to perform an as yet unknown function. Unfortunately, none of the embryonic sections tested contained spleen tissue, such that the expression of *CECR1* shown on the Northern blot (Figure 3-4) could not be confirmed, although CES patients do not show any spleen malformations.

In summary, the RNA *in situ* hybridization results for both pig and human embryos were consistent with CES features, particularly the major heart and kidney defects. Ultimately though, the detection of the CECR1 protein in these tissues is necessary to support and confirm these results, since the presence of the mRNA transcript does not guarantee the production of protein in that tissue, especially if the antisense transcript is present. For example, human *CECR2*, another gene in the CES critical region, is expressed in a wide variety of tissues, but the protein was detected (by gene fusion to β-galactosidase) mainly in the nervous system and limbs (Banting et al., 2005). All of the *in situ* data therefore must be confirmed by Western analysis and protein localization.

Use of the sense probes as negative controls in both the pig and human *in situ* hybridization experiments accomplished two different tasks. For the sense probes located near the 5'end of the gene, an actual negative signal was obtained to serve as a control and validate the experiment. The sense probes derived from the 3' end (pig CECR1-2S, and human CECR1-4S), however, were instrumental in the discovery of a putative antisense transcript (see next section). In hindsight, however, since it is unknown how much of the sense transcript overlaps with the antisense transcript, use of an unrelated positive control (positive signal in different tissues) would have been more useful to substantiate the *in situ* results. For example, the *Pax-1* gene is expressed only in specific somites, depending on the developmental stage under consideration (Barnes et al., 1996), and therefore would not show staining in the liver or kidney tubules.

Discovery of an antisense transcript indicates that *CECR1* may be regulated post-transcriptionally

Proof of existence for the antisense transcript

It has been suggested that as many as 20% of human genes may be influenced by antisense transcripts (Chen et al., 2004). Evidence of a *CECR1* antisense transcript was serendipitously uncovered in the pig *in situ* hybridization studies while using the 3' sense probe (see previous section). This observation was confirmed using the 3'-most sense probe with human slides. In fact, three different experimental procedures (RNA *in situ* hybridization, RT-PCR, and Northern analysis) in two different species (pig and human) all showed the presence of an antisense transcript. Thus, although no ESTs could be identified, and the AntiHunter program failed to predict an antisense transcript from genomic sequence, the presence of an antisense transcript in the *CECR1* region has nonetheless been confirmed beyond doubt. Perhaps the antisense transcript is not terminated with a polyadenylation signal, or is not composed of an extensive open reading frame, which would make it difficult to predict using current computer programs.

Since both transcripts seem to be often expressed in the same temporal and spatial locations, perhaps the antisense transcript is acting to regulate the *CECR1* gene at the post-transcriptional level. It is thought that antisense regulation might be a way of lowering the abundance of stable transcripts more rapidly than the cessation of transcription (reviewed in Chen et al., 2005). Rapid and strict regulation might indeed be required for *CECR1* if it functions as a growth factor and is involved in the development of the various tissues it is expressed in.

The CECR1 *sense and antisense transcripts are differentially expressed*

The relative levels of the sense and antisense transcripts may indicate the amount of CECR1 protein produced. For example, in the day 20 pig using the CECR1-2S sense probe, the antisense transcript was not detected very strongly in the heart compared to the CECR1-1AS probe (Figure 3-13). This suggests that the antisense transcript may not have an effect on the *CECR1* transcript at this stage and might therefore indicate that the CECR1 protein has an important function at this time-point. But in the case of the day 20 pig liver and kidney, the antisense and sense transcript are both expressed (Figure 3-13), suggesting that there might not be any CECR1 protein produced in these tissues. On the other hand, a slight difference in the transcript levels may not be distinguishable using the *in situ* hybridization technique, and therefore protein may be produced at very low levels. The antisense transcript was especially strong in the day 28 pig liver section (Figure 3-14), whereas in the later stage (pig day 31, Figure 3-15) the levels of the antisense transcript were perhaps equal to the sense transcript again, as seen in patches of light staining.

In the day 34 human embryo sections, the liver and kidney tubule staining from the antisense (CECR1-1AS & -4AS) and sense (CECR1-4S) probes seemed to be about equal (Figure 3-16, 3-17, and 3-18), keeping in mind that the CECR1-1AS probe in Figure 3-17 is more heavily stained from a separate experiment. In the day 47 embryo sections, however, the liver staining was dramatically reduced with the CECR1-4S probe compared to the CECR1-1AS probe, and the kidney tubule staining was non-existent (Figure 3-19). Also, the staining with the 4S probe in the epithelial lining of the week 8.5 stomach (Figure 3-22) and 10.7 week liver cells (Figure 3-23) was less intense than with the 4AS probe. The absence/decrease in staining with the CECR1-4S probe at these later stages suggests that the antisense transcript is down-regulated, which may indicate that these stages might be important for *CECR1* expression and function in these tissues.

Type of overlap involved with the antisense transcript

The type of overlap involved in these transcripts might be "tail-to-tail," since the region of overlap was discovered at the 3' but not 5' end of the *CECR1* gene, and since this type is the most common of the antisense overlaps (reviewed in Makalowska et al., 2005). The entire gene structure of the antisense transcript must be determined using 3' and 5' RACE to determine the extent of the overlap. For most sense-antisense pairs, a 5' or 3' UTR is involved in the overlapping region (reviewed in Lehner et al., 2002). Although the extent of overlap for *CECR1* is unknown presently, the length of the 3'UTR (2.2 kb) suggests that it might perform some function, and further might overlap with the antisense transcript due to the probable tail-to-tail type of overlap. It is known that UTRs often play roles in mRNA transport, stability and translation efficiency (reviewed in Lehner et al., 2002), therefore the *CECR1* antisense transcript might be acting to regulate any of these functions for the *CECR1* sense transcript through overlap with the 3' region.

In its location at the 3' end of *CECR1*, the antisense transcript would have an effect on both the variant 1 and variant 2 sense transcripts. The expression of variant 1 or 2 in a specific tissue, therefore, would be dependent on the relative amounts of these two variants, plus the expression of the antisense transcript. It is likely that the antisense transcript would affect both sense transcripts equally, such that the transcript that is present in a lower amount might be completely silenced by the antisense transcript, while the more abundant sense transcript might be translated in a limited capacity. On the other hand, if the extent of overlap extends to the 5' end of either variant, the antisense transcript might cause more of an affect on one sense transcript over the other.

Possible mechanisms involved in regulation of the CECR1 *transcript*

Although the exact mechanisms are not completely understood in humans, numerous studies in eukaryotic cells have shown that antisense transcription can regulate gene expression by three general mechanisms: transcriptional interference, RNA masking, and double-stranded RNA (dsRNA)-dependent mechanisms (Lavorgna et al., 2004a), but these mechanisms may not be mutually exclusive. Although the mechanism of regulation with the *CECR1* antisense transcript is unknown presently, some mechanisms can be ruled out as possibilities based on the evidence obtained thus far.

Transcriptional interference is defined as the direct suppressive influence of one transcriptional process on another (Shearwin et al., 2005). It relies on the bulkiness of the RNA polymerase complex in that only one of the sense/antisense pair can be transcribed at any given time, due to steric hindrance that would prevent two polymerase complexes from passing each other (Lavorgna et al., 2004a). In this way, transcriptional interference is usually linked to the presence of inversely correlated expression patterns (Gibson et al., 2005). Since the *CECR1* sense and antisense transcripts seem to be expressed at the same time in most tissues, transcriptional interference likely is not the mechanism occurring here. The transcription of each transcript, one after another, in the same cell cannot be completely ruled out, however.

The formation of RNA duplexes between sense and antisense transcripts might also cause the "masking" of key regulatory features in either transcript that are necessary for the binding of important *trans*-acting factors (reviewed in Lavorgna et al., 2004a). Some examples of important protein-RNA interactions that might be affected by the presence of an antisense transcript include alternative splicing, mRNA transport, polyadenylation, translation, and degradation (Lavorgna et al., 2004a). RNA masking is a very likely mechanism for the *CECR1* sense-antisense pair, since the presence of both transcripts allows the binding of one to the other, which might prevent certain *trans*-acting factors from binding.

Double-stranded RNA (dsRNA) mechanisms may regulate transcription by either RNA interference (RNAi) or RNA editing mechanisms (reviewed in Gibson et al., 2005). RNA interference involves the cleavage of dsRNA to produce short interfering (si)RNA molecules that form part of the complex responsible for exerting a silencing phenotype (reviewed in Almeida and Allshire, 2005). Silencing can include mRNA degradation, or chromatin modification (methylation) resulting in transcriptional gene silencing (Almeida and Allshire, 2005). RNAi is probably not involved with the *CECR1* region, since it leads to the degradation or prevention of transcription of one or both transcripts, and both the *CECR1* sense and antisense transcripts were detectable by Northern analysis in both pig and human tissues (Figure 3-25 and 3-26).

RNA editing occurs through "adenosine deaminases acting on RNA" (ADARs), which have no sequence similarity to the classic ADAs, but catalyze the hydrolytic deamination of approximately 50% of the adenosines on each strand of the RNA duplex to inosines (reviewed in Lavorgna et al., 2004a). These hyper-edited molecules are retained in the nucleus and thus the translation of both messages is prevented (Lavorgna et al., 2004a). RNA editing has not yet been demonstrated in mammals, therefore it is unknown as to the role that it plays in antisense regulation in humans, but this mechanism might be likely for the *CECR1* transcripts. Since RNA editing occurs on a one-to-one basis, the amount of antisense transcript present would in essence regulate how much sense transcript gets translocated to the cytoplasm to be translated, and vice-versa (Carmichael, 2003). Thus, in situations where the *CECR1* antisense transcript is more abundant, it might attenuate the sense transcript such that no CECR1 protein could be produced. On the other hand, when the amount of *CECR1* message outweighs the number of antisense transcripts, the CECR1 protein could then be produced. In cases where the *CECR1* region is duplicated (in a CES patient), the relative amounts of the sense and antisense transcripts might be altered slightly, which may in turn facilitate some of the features of CES. The detection of the CECR1 protein with an antibody is therefore especially important, to help determine if the antisense transcript is involved in down-regulating translation of the *CECR1* mRNA. However, the presence of the antisense transcript may present a problem if it is indeed preventing the translation of the sense transcript into the CECR1 protein, since there might not be enough protein to detect with the antibody.

There are six *ADGF* homologues in Drosophila

Gene structure and expression pattern individualize the ADGF *genes*

The striking similarities in gene structure, sequence, and intron position, along with the physical clustering in groups within the six Drosophila genes (Figure 3-28 and 3-30) suggests that there may be functional redundancy between the six genes. Based on these properties alone, the six genes can be placed into three groups: *ADGF-B* & *-E*, *ADGF-C* & *-D*, and *ADGF-A* & *-A2*, in that within each group, gene duplication has most probably occurred most recently. The phylogenetic analyses supported these groupings (Figure 3-35). The differing predicted cytological localization and expression patterns, however, individualize each gene such that none are expected to be completely redundant with any other gene.

For example, *ADGF-A* & *-D* are not most similar in genomic structure or sequence to each other, but both proteins have predicted signal peptides (Table 3-1) and have indeed been shown to be secreted into culture medium as growth factors (Zurovec et al., 2002). These two genes also were expressed in all stages tested, when the RT-PCR results are taken into account (Figure 3-31). ADGF-C is also predicted to be secreted, and has a similar expression pattern to ADGF-A & -D, except that no embryonic expression was detected. ADGF-A2, -B, & -E have similar expression patterns, in that they all appear to be male-specific based on the Northern analysis, except that ADGF-E is faintly expressed in females as detected by RT-PCR. Additionally, ADGF-A2 has been shown by *in situ* hybridization to be expressed exclusively in the testes, and is likely a membrane-bound signaling molecule required for spermatogenesis (Matsushita et al., 2000), whereas ADGF-B & -E are predicted to possess mitochondrial targeting peptides. Since there were two different sized transcripts observed on the Northern blots for the *ADGF-A* and *ADGF-E* genes (Figure 3-31), this suggests that alternate splicing and/or alternate polyadenylation signals may be present, adding further complexity to this set of genes.

Therefore, although there is still much similarity within some of the aforementioned three gene groups, none of the six *ADGF* homologues overlaps completely when all of the information is taken into account. Thus it seems that none of the *ADGF* genes is completely redundant with any of the other five genes, and each gene may have evolved different functional roles in Drosophila development and/or metabolism. Ultimately though, any partial redundancy will need to be addressed by looking at mutations in individual genes and in combinations of genes.

Gene orientations and theories of regulation

The close head-to-head orientation of the *ADGF-C & -D* genes suggests that coordinate regulation exists within this gene set. The close proximity of the 5' ends of *ADGF-C & -D* suggests that the promoters may be in the same vicinity, and in fact their Genscan-predicted promoters are separated by only 23 bp. This suggests that promoter competition may occur in order to regulate which of the two genes is expressed. Promoter competition occurs when the RNA polymerase binds one promoter, which prevents the binding of a second RNA polymerase on the promoter in close vicinity (Shearwin et al., 2005). This would prevent both of the transcripts from being expressed at the same time, and is therefore usually associated with reciprocal expression patterns of the overlapping genes (Gibson et al., 2005). The Northern blot analysis for the *ADGF-C & -D* genes showed instead that the expression patterns overlap (Figure 3-31). This does not preclude a finer regulation of expression based on transcriptional interference though, since the Northern data is based on whole animals at different stages, and the two genes could be expressed in non-overlapping tissues at each stage. It might be interesting to determine the expression patterns more specifically, in order to determine if promoter competition is indeed acting to regulate the transcription of the *ADGF-C & -D* genes. This would also aid in the ascertainment of why there are so many *ADGF* genes in Drosophila.

The *ADGF-A2* gene structure is nested within the first intron of *ADGF-A* (Figure 3-28B). The nested arrangement of one gene within an intron of another is quite common, with one estimation suggesting that as many as 7% of Drosophila genes are nested within others (Ashburner et al., 1999), however the two genes are almost always in the opposite orientation (Gibson et al., 2005). Since the *ADGF-A & -A2* genes are structured in tandem, their promoters may be in close vicinity to each other and transcriptional interference might occur with this gene set. RT-PCR in the 75A region revealed that the *ADGF-A2* gene also sometimes shares the 5' exon originally thought to belong to the *ADGF-A* gene (Figure 3-29), suggesting that the two transcripts may rely on the same promoter, and alternate splicing may be involved in regulating their expression. *ADGF-A* is expressed in all stages tested (Figure 3-31) while *ADGF-A2* is specifically expressed in the testes (Matsushita et al., 2000). If alternative splicing plays a role in the decision to express one of these genes over the other, there may be splicing factors present in the testes to promote the splicing of the 5' exon to the rest of the *ADGF-A2* gene. Different splicing factors present in other tissues might instead promote the splicing of the entire first intron of *ADGF-A*, which contains *ADGF-A2*, thus excluding its expression.

Since the utilization of the 5'exon by *ADGF-A2* changes the N-terminus of the protein by 19 amino acids compared to the published sequence (Matsushita et al., 2000), it may represent an internal mechanism of regulation if the two different protein outcomes (with and without the 5' exon) have an effect. This loss of these amino acids does not change the protein localization, since the putative membrane-spanning segment located nearby is not affected, but it may modulate the binding of cofactors for example, which may change the function and downstream effects of ADGF-A2. This would add yet another aspect of complexity to the already very complex and dynamic expression profile of the six *ADGF* homologues.

Theories of gene duplication and divergence

It is interesting that six *ADGF* homologues were discovered in Drosophila, compared to only one human *CECR1* gene, since the converse is usually found. For example, there are eight beta-integrin genes in human, but only two in Drosophila (Schmitt and Brower, 2001). The number of Drosophila genes along with their structure/sequence similarities (Figure 3-28) suggests that at least five duplications along with subsequence divergence of these genes has occurred in the fly lineage. Gene duplication is not uncommon in Drosophila, since it has been suggested that over 5000 genes in the Drosophila genome appear to have arisen by gene duplication to become members of multigene families (Rubin et al., 2000). The fact that the human *CECR1* gene structure also shares two intron locations (Figure 3-30) supports the orthologous nature of these genes.

After gene duplication, one gene might maintain the original function while the other copy is free to accumulate amino acids changes and assume a distinct function or new tissue specificity by chance. For example, while some *D. melanogaster* proteins have been shown to harbour ADA activity (ADGF-A & -D), others, such as ADGF-E, do not (Zurovec et al., 2002). ADGF-E does not share four of the eight conserved ADA residues (Figure 3-32), which may explain its lack of ADA activity. The replacements for these four residues, however, are faithfully conserved between the three Drosophila ADGF-E sequences (*D. yakuba* ADGF-E was not shown), suggesting that this paralogue has evolved a new conserved function. This type of divergence is referred to as neofunctionalization (Lynch and Katju, 2004).

Alternately, the two genes might undergo subfunctionalization such that the all-encompassing function and/or expression pattern of the ancestral gene is lost in a complementary gene-specific manner between the two genes (Dermitzakis and Clark, 2001). The ADGF-A2 gene for example, has a predicted transmembrane domain, suggesting it may have secured a separate functional location compared to the ADGF-B gene, which has a predicted mitochondrial localization signal. Conversely, there may be some redundancy between the six *ADGF* genes, due to the lack of divergence over time.

Possible function of the six Drosophila homologues

Since the six *ADGF* genes have distinctive localization signals and expression patterns, they might have different roles in Drosophila development. The Northern and RT-PCR analyses showed male specific/predominant expression by four of the six genes (Figure 3-31), suggesting a partially redundant theoretical role in male-specific development, although no male sterile mutations were found to be mapped to the chromosomal locations of these genes. Extracellular adenosine levels must be tightly regulated, since different cell types have different adenosine optima (Franco et al., 1997). It has been suggested that the different and highly tissue-specific expression patterns of the six Drosophila ADGFs may reflect the evolution of a mechanism to regulate local extracellular adenosine levels, to provide the appropriate environment for different cell types (Zurovec et al., 2002).

The function of some of the six Drosophila *ADGF* genes has recently been studied using a technique involving loss-of-function (LOF) mutations and gene conversion (Dolezal et al., 2003). LOF of *ADGF-A* caused a larval lethal phenotype, mostly in the late third instar. The larva showed disintegration of the fat body and most individuals developed melanotic tumors (Dolezal et al., 2003). This phenotype makes sense when considering the wide-spread expression profile of *ADGF-A*, since a strong lethal phenotype might be expected from the loss of such a widely expressed protein. This severe phenotype also underscores the fact that no other *ADGF* homologue is able to compensate for the loss of *ADGF-A* and none are therefore completely redundant with this gene. It was further shown that *ADGF-A* expression is specifically required in the blood fluid (hemolymph) of Drosophila larvae, in order to control adenosine levels and therefore the onset of premature metamorphic changes (Dolezal et al., 2005).

Mutants in either *ADGF-C* or *-D* showed lethargy after emerging and semilethality during larval and pupal stages (low penetrance), but the double mutants showed cumulative effects (high penetrance), suggesting that the functions of *ADGF-C & -D* are partially redundant (Dolezal et al., 2003). These results are in agreement with the similar expression patterns and high sequence similarity between these two genes. Mutations in either *ADGF-A2* or *-B*, or the double mutant, did not express any obvious phenotype, and the adults were fertile (Dolezal et al., 2003). This is puzzling, since *ADGF-A2* has been shown to be expressed in mature primary spermatocytes, and was thought to play a role in spermatogenesis (Matsushita et al., 2000). The authors suggested that since *ADGF-E* also shows a similar expression pattern (male predominant), perhaps there is redundancy between all three (ADGF-A2, -B, & -E) genes (Dolezal et al., 2003). Since ADGF-E has been shown to lack ADA activity (Zurovec et al., 2002) due to changes in active site residues, however, it may not have growth factor properties, and may in fact be a pseudogene.

Phylogenetic analysis reveals an evolutionary relationship between the ADGF, ADAL, and ADA subfamilies

Parameters within the Bayesian analysis and differences from Maximum parsimony

There were two aspects of Bayesian analysis that were addressed in the study of the adenyl-deaminase family: the effect of different heating temperatures and the concordance of posterior probability values between runs. Various temperature settings were tested in the initial trial of the ingroup, and the resulting acceptance values for chain swaps were usually not all within the suggested range (10-70%). Increasing the number of generations to 550,000 from 150,000 seemed to increase the number of times that all values fell in the correct range, but it would have been better if every trial received sample trees from each of the four chains. Perhaps if the individual chains had been run longer, the four chains might have had a better chance to converge and the resulting acceptance values would have been in the appropriate range. In fact while this thesis discussion was being written, a MrBayes analysis using 2 million generations was run, which produced the same topology with acceptance values that were all in the correct range. This analysis would have to be repeated multiple times, however, to confirm that the increased number of generations consistently produced appropriate acceptance values.

Another problem that arose between the five 550,000 generation runs was the disagreement in posterior probabilities for three nodes. Most nodes had less than a 5% standard deviation between the five runs, which might be expected from the sampling methodology intrinsic to Bayesian analysis. But the three nodes with a larger standard deviation indicated that certain runs sampled more of one topology versus others. Perhaps if more generations were sampled from, the various topologies would have been sampled more equally between runs, and would therefore stabilize the posterior probabilities across different runs. Again, many runs would need to be completed in order to determine with statistical significance whether the increased generations made a difference. Aside from these two issues, the tree topology resulting from each of the five MrBayes runs was identical, and therefore the conclusions drawn from the analysis are probably credible.

Both the Bayesian and MP trees showed the ADGF, ADAL, ADA and ADE subgroups as separate, well-defined splits that were very well supported by both analyses. The major mismatch between the two methods involving the grouping of the vertebrate ADGFs within the insects in the MP tree was tested by using that topology as a starting tree for a MrBayes analysis. Considering the MP tree was not maintained in the MrBayes run, the lack of internal support within the MP topology, as well as the better fit of the Bayesian tree with the established organismal phylogeny, Bayesian Inference was perhaps a better measure of the phylogeny of this data set.

Conservation of ADA active site residues

As mentioned previously, many ADGF members including *S. peregrina* ADGF-A (Homma et al., 2001), *A. californica* MDGF (Akalal et al., 2003), and *D. melanogaster* ADGF-A and -D have been shown to possess ADA activity that is critical for their mitogenic activity of embryonic insect cells in culture (Zurovec et al., 2002). Indeed, even bovine ADA stimulated the insect cells to proliferate, while the addition of inosine had no effect, suggesting that it is the depletion of adenosine and not the production of inosine that promotes growth (Zurovec et al., 2002). The *L. longipalpis* ADGF protein (Charlab et al., 2001), and salivary extracts from *G. morsitans* (Li and Aksoy, 2000), *C. quinquefasciatus* and *A. aegypti* (Ribeiro et al., 2001) have been shown to possess ADA activity. Interestingly, no ADA activity was found in the salivary glands of *A. gambiae* (Ribeiro et al., 2001), although there have been indications found in this insect of four ADGF homologues. Recently, human CECR1 was also shown to possess ADA activity (Zavialov and Engstrom, 2005).

The crystal structure of mouse ADA has identified amino acid residues with a specific role in the function of the protein (Wilson et al., 1991). Except for some of the insects, all of the ADGF and ADAL members have retained all eight residues required for ADA activity, while the AMPD and ADE families did not conserve all eight (Figure 3-32). The three residues involved in salt-bridge formation, Arg101, Glu260, and Ser265, are also conserved in the ADGF and ADAL subfamilies, but not in the ADEs. Some auxiliary residues thought to be important for ADA function, however, including Asp19 and Ala183 (Wilson et al., 1991), were not conserved within the ADGF or ADAL subfamilies. This suggests that ADGF and ADAL might react with slightly different substrates, or may indicate that different kinetic properties are involved in the reaction, such as with ADA2/CECR1. Both the *L. longipalpis* ADA and the *A. californica* MDGF proteins have been modeled based on the structure of mouse ADA, which showed that all the active site residues in the two ADGF subfamily proteins were conserved in the correct structural locations (Charlab et al., 2001; Akalal et al., 2004). Together with the conservation of all eight ADA active site residues, this indicates that like all ADAs and some ADGFs mentioned above, all the members of the ADGF and ADAL subfamilies may in fact possess ADA activity. Also, the novel MPKG motif is conserved in almost all ADGFs, with the PK being conserved in both the ADA and ADAL subfamilies, suggesting this region of the protein may be important in the overall function, although the significance of this domain is unknown presently.

Interestingly, all the conserved ADA residues are present in *D. melanogaster* ADA (actually a member of the ADAL subfamily and therefore renamed to *D. melanogaster* ADAL; see below) but a recombinant form of this protein did not show ADA activity (Zurovec et al., 2002). This indicates that either *D. melanogaster* ADAL (and perhaps every ADAL member) lacks ADA activity, perhaps due to redundancy with ADGF, or that the recombinant form is not active, or that the correct physiological conditions were not present. If *D. melanogaster* ADAL indeed lacks ADA activity, it may have accumulated mutations that eventually led to the loss of catalytic activity. The duplication and divergence in expression of the six ADGF homologues, the same number of which are not present in other groups of animals, may have taken over the function of ADAL. Once the function of the vertebrate ADGFs and ADALs has been determined, it will be possible to compare the conserved residues within this entire family to assign specific functional roles for each residue.

Since CECR1 variant 2 lacks the MPKG domain and the first two amino acids of the ADA active site (His15 and His17), it is uncertain whether it has ADA or related activity, although it would be expected to lack ADA activity. Interestingly, a testis-specific variant of murine ADA has been found that also lacks the first two amino acids of the active site (Meng et al., 1997), but the ADA activity of this shorter transcript was not tested. The two different *CECR1* transcripts may be performing different functions in the various tissues they are expressed in. For example, without His15 and His17, the CECR1v2 protein may have an effect on a substrate other than adenosine, may bind to a different receptor, or might induce an alternate signal in the cell. Alternately, this alternate splice product may perform a negative regulatory function on the CECR1v1 protein, as discussed in a later section of this chapter.

Patterns revealed in the phylogenetic analyses

Several novel protein sequences with membership in the adenyl-deaminase family were discovered throughout this project, although since most of these new additions were merely predictions, their existence and actual sequence still need to be confirmed. The various phylogenetic analyses revealed that the ADGF and novel ADAL subgroups are clearly related to the classic ADA subfamily. The existence of these three closely related protein subgroups raises the issue of redundancy between ADAL and ADA, and perhaps ADGF. Why have three separate groups of proteins evolved to carry out the same apparently simple function? Although no ADAL subfamily members have been shown to have ADAL activity, if it is proven that they do, there would be three subfamilies, ADAL, ADGF, and the classic ADAs, with members that harbour ADA activity.

The ADAL group is more closely related to the ADA and ADE subgroups, in both sequence similarity and number of residues, compared to the ADGF group. The presence of the ~100 amino acid N-terminal extension in the ADGFs compared to the ADA and ADAL subgroups represents the major size difference. The function of this extension is unknown, but it may be important for substrate specificity, protein-protein interactions, enzymatic activity, or cellular localization, to name just a few possibilities. Besides the size differences, many of the ADGF members were shown to have a predicted signal peptide, whereas none of the ADA or ADAL proteins were, which further confirmed their similarity to each other.

The analysis has also shown that the ADE subfamily arose from the common ancestor with the ADA family. If the vertebrate ADGF and ADAL members are proven to have ADA activity, it is possible that the common ancestor of the entire ingroup had ADA activity, and that only the single-celled organisms gained ADE as a selective advantage. Bacteria lack both an ADAL and ADGF homologue, while fungi lack an

ADA homologue. Since ADE catalyses a slightly different reaction than ADA (Figure 1-6), it may confer a faster turn-over rate of nucleotide substrates than would otherwise occur due to the lack of a full complement of ADA activity. The advantage of retaining ADA activity among three different gene families (ADA, ADAL, and ADGF) in multicellular organisms might have been to compartmentalize the activity, both temporally and spatially. Altogether, it seems that ADA activity is a much more complicated story than previously thought.

Overall, the phylogenetic analysis revealed that some genes previously thought to be classic ADAs are more correctly placed elsewhere. *D. melanogaster* ADA (as named in FlyBase) is a member of the ADALs, while *L. longipalpis* ADA (LuloADA), *C. quinquefasciatus* ADA, and *A. aegypti* ADA belong with the ADGF subfamily. It therefore seems that these insect proteins have been incorrectly labeled. The renaming of *D. melanogaster* ADA to *D. melanogaster* ADAL; *L. longipalpis* ADA to *L. longipalpis* ADGF; *C. quinquefasciatus* ADA to *C. quinquefasciatus* ADGF; and *A. aegypti* ADA to *A. aegypti* ADGF would better reflect their position within the adenyl-deaminase family.

Missing members in different organisms

It seems, therefore, that none of the insects studied have a homologue of the classic ADAs. *D. melanogaster* and *A. gambiae* have complete or nearly finished genomic sequence, and since a classic ADA homologue has not been found in these two organisms, it was probably lost in insects due to the multiple ADGF paralogues that have presumably replaced its function. In fact, there are several different ADGF, ADAL and ADA subfamily members that seem to be missing from completely sequenced genomes. There is no *M. musculus*, *R. norvegicus*, or *C. elegans* homologue in the ADGF subfamily, but homologues of the AMPD, ADA, and ADAL exist in these organisms. This suggests that the ADGF homologue has been lost in these organisms, and perhaps one of the other subfamily members may be compensating for the loss. ADAL was found in insects, vertebrates, and most fungi, but not in prokaryotes. Also, because there were no prokaryotic orthologues of ADGF or ADAL, this suggests that these proteins were gained on the lineage leading to extant eukaryotes. Three fungal ADE members have been discovered in different organisms, but there have been no fungal ADAs found, which could be due to the unfinished state of many fungal genomes, except that one might expect to find an ADA homologue in at least one of the six fungal genomes searched.

Unlike ADGF, there is only one ADA and one ADAL homologue found in all three fish species studied, indicating that either these were not part of the major gene duplication event (Taylor et al., 2003), or that the duplicates of these family members were lost. It is interesting that there is no *S. cerevisiae* homologue of any other subfamily members besides ADE, especially since *S. cerevisiae* ADE has been mistaken previously for a classic ADA in the literature. As mentioned previously, *E. coli* ADE exists but was not included in the phylogenetic analysis, and although both *E. coli* and *S. coelicolor* possess an ADA and ADE gene product, no other family members were found in either bacterial species. All of this data suggests that certain protein subfamilies may be specialized for certain organisms or may procure an advantage in different physiological conditions, or that perhaps the various subfamilies are partially redundant.

For organisms with unfinished genomes, the lack of a certain gene product may be due to a loss of that gene in the organism, or simply that it has not been sequenced yet. This may be especially true for organisms with almost no genomic information, including *D. discoideum, X. laevis, S. scrofa, A. californica, L. longipalpis, C. quinquefasciatus, A. aegypti, S. peregrina*, and *G. morsitans*.

The presence of multiple ADGF members in some organisms suggests exploitation of alternate functions

Many organisms have multiple ADGF paralogues, including the fish, fungi and insects, which may indicate the importance of the ADGF protein for development. It was stated in the results section that within the insect ADGF family, the three Drosophila species acted as a backbone onto which the other insect genes from incomplete genomes may be placed. Since there are already six paralogues within Drosophila, it seems odd that both *A. gambiae* and *G. morsitans* apparently have two genes that are more similar to each other than to the Drosophila orthologues (as seen in Figure 3-35). This suggests that, in addition to the six possible paralogues similar to Drosophila, these two organisms may have separately undergone an additional duplication event to produce a seventh paralogue. Conversely, these results might be explained by gene conversion (reviewed in Papadakis and Patrinos, 1999), such that the two genes appear to be more similar to each other than to one of the other Drosophila homologues. Finally, the duplications observed in the Drosophila species might not have occurred before the divergence of all the insect species, and the one to three genes found in each of the other insect species may have been scattered on the Drosophila backbone simply according to the amount of sequence similarity in these genes. The availability of finished genome sequence for the other insect species may help to clarify this issue.

It is likely that some of the many paralogues found in insects have acquired a specialized expression pattern, and may have adopted a broader range of functions. Since ADGF expression has been observed mainly in the salivary glands of biting insects (Li and Aksoy, 2000; Charlab et al., 2000), it has been suggested to aid insects in providing pain relief at the site of biting (Charlab et al., 2001). In other words, perhaps the *ADGF* genes have been adapted for a specialized physiological purpose in some insects. The presence of ADA activity in hematophagous (blood-sucking) insects is contradictory to the positive effects that adenosine has on vasodilation and platelet aggregation inhibition that would be expected to increase blood availability while feeding (Ribeiro et al., 2001). But the presence of adenosine at the bite site also elicits a negative effect; the induction of histamine release that causes itching at the feeding site, and alerts the host to the insect's presence. This negative effect may be relieved by ADA activity at the feeding site, making the presence of ADA activity advantageous over its absence (Ribeiro et al., 2001). Also, inosine produced by ADA activity has been shown to inhibit the production of inflammatory cytokines (Hasko et al., 2000), which may further aid in insect concealment. This function of ADGF must have evolved specifically for biting insects, since neither Drosophila nor *S. peregrina* are biting flies.

Another aspect of ADA activity in conjunction with biting insects is the transmission of parasites through the insect's mouthparts. *G. morsitans* transmits various types of the trypanosome parasite, which cause human African sleeping sickness, and the expression of ADA activity in the salivary glands may be amenable to parasite survival (reviewed in Li and Aksoy, 2000). The TSGF proteins may be involved in affecting the maturation process of parasites, and adenosine deaminase activity might modulate the host immune response in the presence of infectious parasites (Li and Aksoy, 2000), although it is not clear what selective advantage this holds for the fly host. Perhaps there is a symbiotic relationship between the trypanosome parasite and *G. morsitans* that has yet to be discovered. Since no ADA activity was found in the salivary glands of *A. gambiae* (Ribeiro et al., 2001), perhaps the four ADGF paralogues in this insect have taken on a different albeit unknown function as well.

Therefore in general, some of the functions of the ADGFs in insects may involve growth of embryonic cells (Homma et al., 1996; Zurovec et al., 2002), pain relief at the bite site (Ribeiro et al., 2001; Charlab et al., 2001), and maturation of parasites (Li and Aksoy, 2000), among other functions not yet uncovered. Therefore, further characterization of ADGF genes in insects may identify more putative functions, some of which might be applicable to vertebrates, while others may be specific to biting flies and have implications for the transmission of fly-born diseases in human health.

Proof for the introns-late aspect of the new synthetic theory of introns

If the intron position data is available, large gene families are useful for studying the relationship of evolution and the conservation of intron positions. Whether spliceosomal introns were present in primitive coding sequences (introns-early) or added later in the lineage leading to eukaryotes (introns-late), or a combination of both (synthetic theory) is an intense area of debate (Gilbert et al., 1986; Fedorova and Fedorov, 2003; de Souza, 2003). The ADA subfamily has both eukaryotic and prokaryotic members, and would therefore be considered ancient. The duplication and divergence of the other three groups might also be considered to be ancient, especially when considering the fungal genes found in each. But only one intron position was conserved between two of the four ingroup subfamilies when intron-sliding was not considered, and that intron position was found in only one member of each of those subfamilies, suggesting it might be a coincidence. If the introns-early aspect of the synthetic theory was to be accepted, it might be expected that more intron positions would be retained between the four subgroups of the ingroup. There were many instances where an ancestral intron might have existed when intron-sliding was taken into account. For each case, however, the most parsimonious reconstruction favoured a few instances of intron gain rather than many more losses. A trend seemed to emerge within the organisms in this study that there are generally more introns found in higher organisms, compared with the lower deuterostomes, which have an intermediate number of introns, and single-celled organisms with no introns. This suggests in general that introns were added along the eukaryotic lineage over time.

This data, although it seems to substantiate the introns-late side, has not entirely ruled out the introns-early aspect of the synthetic theory. Indeed, analysis of the intron phases within the entire ingroup showed a slight excess of phase 0 introns (46%, 145/316). Also, use of the lack of conserved introns to support the introns-late side assumes that intron loss and gain are equally likely. If intron loss was entirely easier that intron gain, the introns-early side may hold some weight with this data. Therefore, it is possible that the resolution to the debate cannot be undertaken until the relative costs of intron loss/gain are determined (Tyshenko and Walker, 1997). In a study of human, coral, fly and worm integrin-β genes, the coral gene shared 25 of 26 intron positions with at least one other species, when intron-sliding was taken into account (Schmitt and Brower, 2001). Without the coral sequence, only 8 splice sites were shared between two or more phyla. This suggests that without an ancestral sequence such as this coral sequence, the results might incorrectly appear to only support the introns-late aspect of the new synthetic theory. Therefore, although the results at present seem to support the introns-late side, the addition of more data as it becomes available may change this view.

Since *CECR1* is missing in mouse, is *ADAL* or *ADA* providing compensation?

When *ADA* is disrupted in mice, the fetuses die perinatally due to severe liver damage (Wakamiya et al., 1995; Migchielsen et al., 1995). When the placenta of *ADA* -/- mice is engineered to produce ADA, however, the pups survive and exhibit metabolic and immunologic features similar to those seen in ADA deficient humans (Blackburn et al., 1998). It is not known whether the *ADA* transcript is normally expressed in placenta, but human *CECR1* is expressed in this tissue (Figure 1-3). Since mice lack a *CECR1* homologue, the results with the *ADA* -/- mice suggest that the ADA activity resulting from *CECR1* expression may be required in the placenta, without which the tissue cannot support the growth of the embryo. Besides these immunologic findings though, ADA deficient mice exhibit severe pulmonary insufficiency, bone abnormalities, and kidney pathogenesis (Blackburn et al., 1998), a phenotype much worse than is found in humans. Given that *ADA* -/- mice have only the function of the ADAL protein remaining, the reason why they are more seriously affected compared to ADA deficient humans might be that both the CECR1 and ADAL proteins are intact in humans. Perhaps both the *ADA* and *ADAL* homologues in mice are compensating for the loss of the *CECR1* gene product, and the loss of one (*ADA*) of the two remaining members of the adenyl-deaminase family is catastrophic. The fact that *CECR1* is predicted to be secreted whereas *ADA* and *ADAL* are expected to be intracellular may not be a factor, considering that ADA-deficient humans are routinely treated with PEG-ADA, which is known not to be transported into the cell efficiently (Hershfield, 1995). It might be interesting to cross *ADA* -/- mice with the human *CECR1* transgenic mice created within this project, to confirm that *CECR1* is able to lessen the severe mouse phenotype.

The fact that mice normally harbour only two of three adenyl-deaminase family members suggests that the expression of all three may partially overlap. The comparison of the expression patterns of all three genes is required to determine if each is able to compensate for the other. It is currently under investigation as to where *ADAL* is expressed (Nic Fairbridge, in progress). Also, ADA and ADAL have predicted molecular weights of 40.8 kDa and 40.3 kDa, respectively (see Table 4-1). Since their sizes are so similar, ADA activity previously attributed to ADA1, the 41 kDa form of ADA (Van der Weyden and Kelley, 1976; Ungerer et al., 1992), might indeed actually result from one or the other, or both. Their similarity in size may have concealed ADAL as a contributor of ADA activity until now. For this reason, it will be important to determine if ADAL has ADA activity as well.

Human *CECR1* as a candidate for cat eye syndrome

Many different tissues and organ systems are affected in CES patients, including the eyes, heart, kidney, ears, face and urogenital area (Schinzel et al., 1981; Rosias et al., 2001). Of the fourteen putative genes discovered in the CES critical region, it is not known which gene(s) cause the phenotype when overexpressed (Footz et al., 2001). It is possible for a multi-systemic disorder to be caused by only one gene. For example, Alagille syndrome, which is characterized by liver failure, heart defects, skeletal malformations, ophthalmological abnormalities and a characteristic facial appearance, is caused by mutations in the *Jagged1* gene (Li et al., 1997; Oda et al., 1997). Alternatively, multi-system disorders also exist where multiple genes must be affected to produce the entire phenotype. Deletion or mutation of the *TBX1* gene causes almost all the phenotypes observed in the 22q11.2 deletion syndrome, except the learning difficulties (Yagi et al., 2003), suggesting that more that this one gene is involved. Multiple genes are also involved in Williams-Beuren syndrome, which is due to a chromosome 7q11 deletion, and characterized by growth retardation, heart abnormalities (due to deletion or mutation in the *elastin* gene), hypercalcemia, cognitive disabilities and facial abnormalities (reviewed in Tassabehji, 2003). Other than *elastin*, none of the remaining twenty-three genes in the Williams-Beuren syndrome critical region have been linked to any part of the phenotype (Tassabehji, 2003). At the outset therefore, since *CECR1* is only one of 14 genes in the CES critical region, it is equally probable that *CECR1* has nothing, something, or everything to do with the features of CES when overexpressed.

The information gathered previous to and within the body of this thesis suggests that *CECR1* might have at least something to do with the features of CES when overexpressed. Northern analysis (Figure 3-4) and RNA *in situ* hybridization studies (Figure 3-16, etc.) showed that *CECR1* is expressed in fetal tissues affected in CES patients, including heart and kidney, among others. Therefore, *CECR1* may be involved in the production of heart and kidney malformations when overexpressed. *CECR1* is not expressed very strongly in the brain, however, and therefore is probably not involved in the mental development of CES patients. Another gene in the CES critical region, *CECR2*, is highly expressed in neural tissue (Banting et al., 2005), suggesting that it may be responsible for the retardation of mental development in CES, and further showing that *CECR1* is probably not involved in mental development.

CECR1 may also be a less likely candidate for the production of coloboma in the eyes of CES patients. There is no *CECR1* homologue in mouse, but there was a remnant found on mouse chromosome 6, which is the syntenic region of the CES critical region (Footz et al., 2001). It is known that mice that are trisomic for chromosome 6 have eye defects, including coloboma, among other abnormalities unrelated to CES (reviewed in Hernandez and Fisher, 1999). This suggests that another gene on chromosome 6 causes coloboma in mice when duplicated, and therefore suggests that *CECR1* is not involved in this eye defect in CES patients. This suggestion is further supported by the *in situ* hybridization data that showed no *CECR1* expression in eye (Figure 3-16). Also, *CECR2* is actually expressed in the developing eye (Banting et al., 2005), suggesting that it is most probably responsible for the eye phenotype in CES patients.

The phylogenetic analysis has revealed the extent of conservation of CECR1 homologues in a wide variety of organisms and therefore suggests that this protein may be important for development. Based on its identity as ADA2 (Zavialov and Engstrom, 2005), and its homology to the ADGF subfamily, human CECR1 may function as a secreted growth factor in human development. The abnormalities observed in CES might be due to changes in the regulation of extracellular adenosine in tissues where the CECR1 protein is expressed, which might cause problems in the growth of those tissues. Therefore, the determination of whether CECR1 has growth factor properties is critical to determining its role in CES.

The involvement of *CECR1* variant 2 in CES

An alternatively spliced *CECR1* mRNA transcript, called *CECR1* variant 2 (*CECR1v2*), was discovered through searches of the human EST database, and starts within intron 3 of *CECR1v1*. The use of an alternative transcription initiation exon is a rare type of alternative splicing (Ast, 2004), and suggests that an alternate promoter site may be present (Landry et al., 2003). Variant 2 was shown by Northern analysis (Figure 3-4) to be expressed in adult heart and kidney, and all fetal tissues tested, suggesting that the *CECR1v2* promoter is specific to these tissues. This expression pattern fits well with the heart and kidney defects observed in CES patients (Schinzel et al., 1981), suggesting that *CECR1v2* might actually be the transcript involved in the production of CES features, rather than *CECR1v1*.

The presence of the *CECR1v2* alternate exon was only found in human, chimp, and baboon (based on blast searches of genomic sequence), and not in more diverged model organisms such as chicken and Xenopus. These two organisms therefore might not be useful to further study the *CECR1v2* gene product and its role in CES. By default, since there is no *CECR1* homologue in mouse or rat, there could not possibly be a *CECR1v2* gene either. The presence/absence of *CECR1v2* in pig could not be determined due to the lack of genomic data, although there were multiple bands on the Northern blot (Figure 3-25B). Comparative analysis between human and mouse orthologous genes revealed that alternative splicing is often associated with recent exon creation and/or loss (Modrek and Lee, 2003), such that alternative splicing has the potential of creating species-specific cassette exons. To determine if *CECR1v2* is either mammal or primate-specific, pig and other mammals must be tested for the presence of the *CECR1v2* gene, perhaps by low stringency Southern analysis using a human probe, or with 5' RACE. If *CECR1v2* is determined to be primate-specific, tissue culture using primate or human cell lines might be the best course of action for studies of CECR1v2 function.

The CECR1v2 protein excludes the MPKG motif and first two ADA catalytic residues (His15 and His17) found in all other ADGF proteins, suggesting that it may not have the same function as CECR1v1 and the other ADGF proteins. Indeed, CECR1v2 may perform an entirely different tissue-specific function, or may interact with CECR1v1 in a dominant negative fashion to serve a regulatory role. For example, the interferon regulatory factor-3 (IRF-3) protein up-regulates the induction of interferon β (IFNβ) by binding to its promoter (Karpova et al., 2000). An alternative splice form that is expressed most highly in brain, called *IRF-3a*, which splices from exon I to an alternative exon IIa located in intron II of the *IRF-3* gene, excludes exon II and therefore most of the DNA binding domain. Without the capacity to bind the *IFNβ* promoter, this splice product may play a protective role in the brain by suppressing the transcription and thus the toxic effect of IFNβ (Karpova et al., 2001). Further, since IRF-3a was shown by coimmunoprecipitation experiments to form a heterodimer with IRF-3, it might prevent the binding of IRF-3 to the *IFNβ* promoter in a dominant-negative manner to therefore prevent IFNβ production (Karpova et al., 2001).

Since CECR1v1 functions as a homodimer (Zavialov and Engstrom, 2005), in tissues where both variants are expressed the CECR1v2 protein might combine with CECR1v1 to produce a heterodimer instead, which may have a different function or may serve to block the function of CECR1v1. This effect might necessarily occur inside the cell, since the CECR1v2 protein is not predicted to be secreted like CECR1v1, but since ADA is in some way released from the cell as ecto-ADA, this stipulation may not hold. If abrogation of CECR1v1 is indeed the function of CECR1v2, then its expression in developing tissues such as fetal kidney may prevent the function of CECR1v1 in these tissues to perhaps allow normal development. Overexpression of the CECR1 region in CES patients might change the overall ratio of CECR1v1 to CECR1v2, depending on the natural ratio, and might therefore adversely affect tissue development. Another possibility is that CECR1v2 binding to CECR1v1 may modulate its function by allowing the heterodimer to bind a certain protein or cellular structure that is not possible when CECR1v2 is not present. This will depend on the co-localization of both proteins, however. Clearly though, the discernment of CECR1v2 function will be very important in determining its role in CES.

CECR1 quite likely therefore has a dual function, in both the development of tissues important in CES, such as heart and kidney, and in the immune system as a defense against intracellular pathogens. The choice of which protein variant to produce, CECR1v1 or CECR1v2, may modulate the functional effect. Perhaps CECR1v2 is required for the developmental aspect, since it is expressed more strongly in the fetal tissues affected in CES, whereas CECR1v1 plays a role in the immune system aspects. Various types of regulation might exist for these *CECR1* genes, including possible dominant negative protein interactions of CECR1v2 against CECR1v1, and antisense regulation of both genes, although these mechanisms need to be proven. Clearly, human *CECR1* and the *ADGF* gene family is an interesting and important system that requires much more research to determine its role in CES, development, and biology in general. Therefore, although it appears that *CECR1* may be a good candidate for involvement in cat eye syndrome, until its function(s) can be confirmed it remains unknown as to the role *CECR1* plays in the production of CES features.

Conclusions

The data collected throughout this project has advanced the characterization of the *CECR1* gene, and has identified *CECR1* as a good candidate for the production of CES features when overexpressed. *CECR1* is expressed strongly in heart and kidney, two tissues important in CES patients. The expression of human *CECR1* is highly complex, however, since two different variants may regulate its functional aspects in various tissues, and antisense regulation may be in place to coordinate the functional aspects in a tissue and/or time dependant manner.

ADA activity is clearly not as straightforward as once thought. The conservation of ADGF sequences in diverse phyla indicates that the ancestral gene function(s) must have been strongly selected to be retained as divergence occurred. The structure and conserved residues of the ADGF and ADAL subfamilies, combined with their evolutionary relationship to classic ADAs, suggests that these three proteins are all involved in ADA activity. If the expression of each is found in a variety of cellular locations, together they may control adenosine levels in a concerted fashion.

It is not known currently whether *CECR1* is involved in the production of CES features. Its expression in various tissues affected in CES, and its strict conservation throughout evolution, however, suggest that it is an important secreted growth factor that might be relevant to CES. Undoubtedly, much more work is required to determine the extent of involvement of *CECR1* in the features of cat eye syndrome.

Future Directions

Use of model organisms for deletion/duplication studies

Ultimately, the discovery of a patient with a microduplication encompassing *CECR1* would vastly help in determining the participation of *CECR1* in the production of CES features. In the mean time, the strengths of certain model organisms can be taken advantage of in order to explore the characteristics of *CECR1*. The search for *CECR1* homologues for use in the phylogenetic analysis revealed many new genes from various organisms. Although CES is caused by a duplication, the deletion of a gene can often provide clues to its normal function. It has been shown that deletion of the various Drosophila *ADGF* genes causes problems in development (Dolezal et al., 2003). *ADGF-A* homozygotes are lethal in the late third instar larva stage, and deficiencies in either *ADGF-C* or *-D* exhibit larval/pupal lethality at a low penetrance separately, and at a higher penetrance for the double mutant (Dolezal et al., 2003). This suggests that ADGF proteins are essential for development. It would be helpful to examine the loss of *CECR1* in other organisms more closely related to humans.

Since there is no mouse *CECR1* homologue, its deletion/overexpression cannot be tested in this model organism. Therefore, other model organisms must be utilized. The creation of specific gene "knock-downs" in zebrafish has recently become available due to the use of antisense oligonucleotides called morpholinos (Nasevicius and Ekker, 2000). Fang Yang is currently testing morpholinos against both zebrafish *CECR1* homologues, in conjunction with overexpression of each respective mRNA, to determine if changing the transcript levels affect zebrafish development.

Chicken embryos can be used for overexpression studies by implanting protein-coated beads within the developing embryo (Watkins et al., 1998). The chicken CECR1 protein could be expressed recombinantly and then coated onto beads to be implanted next to the embryonic heart, to see if this overexpression would cause defects related to CES. Alternately, transgenic chickens could be produced using retroviral vector-mediated transmission of the chicken CECR1 gene into an embryo (Mozdziak and Petitte, 2004).

With the Xenopus *CECR1* gene now available, the genomic segment containing the entire Xenopus *CECR1* gene could be isolated from a BAC library for use in the creation of transgenic Xenopus embryos in order to study the effects of overexpression. Transgenic Xenopus expressing green fluorescent protein (GFP) under the control of a brain-specific promoter have recently been creating, demonstrating the feasibility of this technique (Kelly et al., 2005). The effectiveness of the Xenopus and chicken studies, however, may depend on whether variant 2 is present or not.

Biochemical properties of CECR1

As presented in the introduction, many of the ADGF members have been shown to act as growth factors in tissue culture, a function that is dependent on ADA activity. It must also be confirmed whether CECR1 acts as a growth factor that is dependent on its ADA activity. In order to test this, a human cell line expressing CECR1 could be made, using a mammalian expression vector, and compared to control cells transfected with an empty vector. This experiment is currently underway (Rezika Zurch and Twila Yobb). If a cell line expressing CECR1 can be established, the effects of CECR1 overexpression on adenosine levels and cellular growth rate could be tested, and perhaps give an indication of how CECR1 acts to modulate growth in human tissues. Alternately, native CECR1 could be purified from human plasma using the same biochemical methods used previously (Zavialov and Engstrom, 2005) and added to cells in culture to observe these effects. This purified protein might also be a good antigen for use in making a CECR1 antibody in rabbits, since recombinant methods have failed thus far.

Native PAGE should be completed to assess whether CECR1 exists as a homodimer, like its ADGF-A homologue in *S. peregrina* (Homma et al., 1996), and as suggested through studies of ADA2 (Iwaki-Egawa et al., 2004). Obtaining the crystal structure of the CECR1 protein would also help to determine its function, since the placement of the ADA active site residues within the CECR1 structure could be compared directly to the structure of ADA. Andrey Zavialov is currently attempting to crystallize recombinant human CECR1 in collaboration with the McDermid lab.

Protein localization and binding assays

A good antibody against CECR1 would be useful for studying many additional aspects of this exciting protein. If CECR1 is indeed a secreted growth factor, as suggested by the presence of a signal peptide and homology to other growth factors, a receptor or other interacting protein may exist that is necessary for eliciting a response signal or to otherwise carry out its function. Indeed, the phylogenetically related ADA protein exerts some of its effects by binding to the A_1R receptor or CD26 protein (Franco et al., 1997). Also, the *S. peregrina* ADGF-A protein was shown to bind the cell surface of NIH-Sape-4 cells (Homma et al., 2001). A CECR1 antibody could be used to determine the sub-cellular location of CECR1 by immunofluorescence or immunohistochemistry in order to confirm that the protein is secreted. Interacting proteins of CECR1 could be identified by coimmunoprecipitation. Alternately, the sub-cellular localization and binding partners could be identified using a CECR1 fusion protein to GST (glutathione-S-transferase) and/or GFP. This would avoid the need for a CECR1-specific antibody. On the other hand, since the CECR1 protein may be present in low amounts in the cell, and since recombinant CECR1 may not be functional and therefore not able to bind its putative partner, the above mentioned experiments may be futile. Specific interactions with proteins such as CD26 or the adenosine receptors might therefore be better tested using a technique such as the yeast two-hybrid system (Fields and Song, 1989; Young, 1998), since this *in vitro* method does not depend on the physiological levels of proteins.

Increased confidence in the phylogenetic analysis with more protein sequences

In a few years time, the phylogenetic analysis could be repeated, in order to include the multitude of genes from various organisms that will be available at that time. The discrepancy between the results obtained from MrBayes versus MP will be closer to being resolved with more sequences. Along with more sequences, phylogenetic inference will also be more advanced in the near future. Indeed, during the writing of this thesis a new version of MrBayes was released (version 3.1), which allows the inference of ancestral states, among other new features (http://mrbayes.csit.fsu.edu/index.php).

Further characterization of CECR1v2

CECR1v2 was shown by Northern analysis to be expressed strongly in adult heart and kidney, as well as fetal heart, kidney, and lung, whereas *CECR1v1* was expressed in a variety of tissues (Figure 3-4). Since Northern analysis is not very sensitive, real-time RT-PCR could be used with primers spanning the two different splice junctions. This will more precisely determine the expression pattern of variant 1 versus variant 2 to obtain a complete picture of the alternative splicing events that may be occurring, and to identify the tissues that may harbour both proteins. RNA *in situ* hybridization using variant-specific probes would also be useful in narrowing down the expression pattern of each *CECR1* gene.

Since there has been no upstream stop in the *CECR1v2* ESTs found thus far, 5' RACE (Rapid Amplification of cDNA Ends) should be used in order to obtain more upstream sequence. This procedure may also reveal more of the gene structure of *CECR1v2* and therefore aid in the understanding of the choice between the two alternate forms of *CECR1*. A preliminary 5' RACE experiment by Fang Yang (unpublished data) has recently revealed the presence of an upstream stop in the *CECR1v2* transcript, confirming that the coding sequence as it is currently known is most probably correct. This experiment will need to be repeated to confirm this result and hopefully obtain more of the upstream sequence to justify the transcript size observed on the Northern blot (Figure 3-4).

The production of an antibody specific to CECR1v2 would also be helpful to further characterize the differences between variant 1 and 2. This could be accomplished with a peptide antibody against the ten amino acids present in the alternative exon, but this may be difficult, as short peptides against variant 1 have not worked well in the past. CECR1v2 might perform a regulatory function by competing with the CECR1v1 protein for binding sites or substrates. If a CECR1v2 antibody was successfully produced, it could be used in conjunction with a CECR1v1 antibody to determine if the proteins are co-expressed, and if they are localized together in the cell. If both conditions hold true, and the ADA activity of recombinant CECR1v1 can be established, an assay could be designed to determine if the addition of CECR1v2 blocks the activity of CECR1v1, which might indicate that it is acting in a dominant negative manner.

CECR1v2 was not found in the chicken or Xenopus genome sequences by blast searches, which suggested that it might be primate or at least mammal-specific. The use of model organisms to characterize *CECR1v2* might therefore depend on whether *CECR1v2* is discovered in the currently unfinished genomes of pig or cow. All of these organisms should therefore be tested using 5' RACE. If *CECR1v2* is discovered in the pig genome for example, the overexpression of *CECR1v2* could be studied by the construction of transgenic pigs, although the cost of this endeavor may be prohibitive.

Characterization of the CECR1 *antisense transcript*

Three different methods in two separate organisms confirmed the presence of an antisense transcript to *CECR1* involving its 3' end. To obtain the full sequence and allow further characterization of the antisense transcript in humans, 5' RACE should be completed. RT-PCR could also be employed, to compliment and confirm the RACE results, or as an alternative method if the RACE experiment proves difficult. Although there are currently no ESTs deposited in the database that represent the *CECR1* antisense transcript, some may come available in the future, so periodic database searches should be performed. Also, more advanced software for predicting antisense transcripts may become available, which would make the search easier.

Once the entire sequence is known, the determination of the effect of the antisense transcript on the *CECR1* sense transcript can then begin. The presence of both transcripts in the same temporal and spatial pattern suggests the possibility of antisense regulation, perhaps by either RNA masking or RNA editing. If the antisense transcript extends to the transcription start site of either *CECR1v1* or *CECR1v2*, it might act to modulate which splice form is produced. To rule out the possibility that the *CECR1* gene is being silenced by RNAi, the methylation status of *CECR1* could be determined using methylation-sensitive restriction enzymes at different developmental time points using pig embryos. In order to determine the relative levels of each transcript if RNA editing is involved, real-time PCR could be used for mutually exclusive regions of each transcript in order to detect any small differences in expression levels. Concurrently, the antisense transcript could be co-expressed in a human cell line expressing the *CECR1* sense transcript to determine if protein levels are reduced when the antisense transcript is present. Of course this experiment relies on the availability of a good antibody to CECR1, and the confirmation of CECR1 expression in that cell line; two requirements that may not be easy to attain. As such, this experiment might be better accomplished using a recombinant (tagged) form of CECR1 expressed in the cell line.

Significance of this work

The study of the expression and function of *CECR1* will help to advance the characterization of the ADGF family of growth factors that has been discovered as a result of homology searches with human CECR1. The results of this work will also shed light on the involvement of *CECR1* in the production of the CES phenotype. Although CES is a rare genetic disease, the birth defects associated with it are not. The study of the overexpression of the genes involved in CES will directly apply to the syndrome, provide insight into the cause of various birth defects, and advance our understanding of normal human development and the effect of changes in gene expression on that development.

Table 4-1. Predicted molecular weight (mw) of human proteins in the adenyl-deamiase family.

Protein	Predicted mw
CECR1 isoform a precursor	58.9 kDa
CECR1 isoform a	55.9 kDa
CECR1 isoform b	30.7 kDa
ADAL	40.3 kDa
ADA1	40.8 kDa
ADA2	114 kDa *

The molecular weight of all proteins were predicted using the Expasy Molecular Biology Server's Compute pI/Mw tool (http://au.expasy.org/tools/pi_tool.html), except ADA2 (*), which was obtained from Van der Weyden and Kelley, 1976.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.H., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor, G.L., Abril, J.F., Agbayani, A., An, H.J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., and Bouck, J. (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Akalal, D.B., Bottenstein, J.E., Lee, S.H., Han, J.H., Chang, D.J., Kaang, B.K., and Nagle, G.T. (2003). Aplysia mollusk-derived growth factor is a mitogen with adenosine deaminase activity and is expressed in the developing central nervous system. Brain Res Mol Brain Res *117*, 228-236.

Akalal, D.B. and Nagle, G.T. (2001). Mollusk-derived growth factor: cloning and developmental expression in the central nervous system and reproductive tract of Aplysia. Brain Res Mol Brain Res *91*, 163-168.

Akalal, D.B., Schein, C.H., and Nagle, G.T. (2004). Mollusk-derived growth factor and the new subfamily of adenosine deaminase-related growth factors. Curr Pharm Des *10*, 3893-3900.

Almeida, R. and Allshire, R.C. (2005). RNA silencing and genome regulation. Trends Cell Biol *15*, 251-258.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Andreasyan, N.A., Hairapetyan, H.L., Sargisova, Y.G., and Mardanyan, S.S. (2005). ADA2 isoform of adenosine deaminase from pleural fluid. FEBS Lett *579*, 643-647.

Ashburner, M., Misra, S., Roote, J., Lewis, S.E., Blazej, R., Davis, T., Doyle, C., Galle, R., George, R., Harris, N., Hartzell, G., Harvey, D., Hong, L., Houston, K., Hoskins, R., Johnson, G., Martin, C., Moshrefi, A., Palazzolo, M., Reese, M.G., Spradling, A., Tsang, G., Wan, K., Whitelaw, K., and Celniker, S. (1999). An exploration of the sequence of a 2.9-Mb region of the genome of Drosophila melanogaster: the Adh region. Genetics *153*, 179-219.

Ast, G. (2004). How did alternative splicing evolve? Nat Rev Genet *5*, 773-782.

Baldini, A. (2003). DiGeorge's syndrome: a gene at last. Lancet *362*, 1342-1343.

Banting, G.S., Barak, O., Ames, T.M., Burnham, A.C., Kardel, M.D., Cooch, N.S., Davidson, C.E., Godbout, R., McDermid, H.E., and Shiekhattar, R. (2005). CECR2, a protein involved in neurulation, forms a novel chromatin remodeling complex with SNF2L. Hum Mol Genet *14*, 513-524.

Barnes, G.L., Hsu, C.W., Mariani, B.D., and Tuan, R.S. (1996). Chicken Pax-1 gene: structure and expression during embryonic somite development. Differentiation *61*, 13-23.

Bartsch, O., Rasi, S., Hoffmann, K., and Blin, N. (2005). FISH of supernumerary marker chromosomes (SMCs) identifies six diagnostically relevant intervals on chromosome 22q and a novel type of bisatellited SMC(22). Eur J Hum Genet *13*, 592-598.

Baud, V., Mears, A.J., Lamour, V., Scamps, C., Duncan, A.M., McDermid, H.E., and Lipinski, M. (1994). The E subunit of vacuolar H(+)-ATPase localizes close to the centromere on human chromosome 22. Hum Mol Genet *3*, 335-339.

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M., and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. Genome Res *10*, 1001-1010.

Becerra, A. and Lazcano, A. (1998). The role of gene duplication in the evolution of purine nucleotide salvage pathways. Orig Life Evol Biosph *28*, 539-553.

Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. J Mol Biol *340*, 783-795.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2004). GenBank: update. Nucleic Acids Res *32 Database issue*, D23-D26.

Betts, M.J., Guigo, R., Agarwal, P., and Russell, R.B. (2001). Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution? EMBO J *20*, 5354-5360.

Birchler, J.A., Riddle, N.C., Auger, D.L., and Veitia, R.A. (2005). Dosage balance in gene regulation: biological implications. Trends Genet *21*, 219-226.

Blackburn, M.R., Datta, S.K., and Kellems, R.E. (1998). Adenosine deaminase-deficient mice generated using a two-stage genetic engineering strategy exhibit a combined immunodeficiency. J Biol Chem *273*, 5093-5100.

Braissant, O, and Wahli, W. (1998). A simplified *in situ* hybridization protocol using non-radioactively labeled probes to detect abundant and rare mRNAs on tissue sections. Biochemica 1, 10-16.

Bridgland, L., Footz, T.K., Kardel, M.D., Riazi, M.A., and McDermid, H.E. (2003). Three duplicons form a novel chimeric transcription unit in the pericentromeric region of chromosome 22q11. Hum Genet *112*, 57-61.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol *268*, 78-94.

Carmichael, G.G. (2003). Antisense starts making more sense. Nat Biotechnol *21*, 371-372.

Chang, Z.Y., Nygaard, P., Chinault, A.C., and Kellems, R.E. (1991). Deduced amino acid sequence of Escherichia coli adenosine deaminase reveals evolutionarily conserved amino acid residues: implications for catalytic function. Biochemistry *30*, 2273-2280.

Charlab, R., Rowton, E.D., and Ribeiro, J.M. (2000). The salivary adenosine deaminase from the sand fly Lutzomyia longipalpis. Exp Parasitol *95*, 45-53.

Charlab, R., Valenzuela, J.G., Andersen, J., and Ribeiro, J.M. (2001). The invertebrate growth factor/CECR1 subfamily of adenosine deaminase proteins. Gene *267*, 13-22.

Chen, J., Sun, M., Hurst, L.D., Carmichael, G.G., and Rowley, J.D. (2005). Human antisense genes have unusually short introns: evidence for selection for rapid transcription. Trends Genet *21*, 203-207.

Chen, J., Sun, M., Kent, W.J., Huang, X., Xie, H., Wang, W., Zhou, G., Shi, R.Z., and Rowley, J.D. (2004). Over 20% of human transcripts might form sense-antisense pairs. Nucleic Acids Res *32*, 4812-4820.

Cheng, J. and Grande, J.P. (2002). Transforming growth factor-beta signal transduction and progressive renal disease. Exp Biol Med *227*, 943-956.

Chrast, R., Scott, H.S., and Antonarakis, S.E. (1999). Linearization and purification of BAC DNA for the development of transgenic mice. Transgenic Res *8*, 147-150.

Clarke, E.J. and Allan, V. (2002). Intermediate filaments: vimentin moves in. Curr Biol *12*, R596-R598.

Cordero, O.J., Salgado, F.J., Fernandez-Alonso, C.M., Herrera, C., Lluis, C., Franco, R., and Nogueira, M. (2001). Cytokines regulate membrane adenosine deaminase on human activated lymphocytes. J Leukoc Biol *70*, 920-930.

Cormack, D.H. (1993). Essential Histology. Philadelphia: J.B. Lippincott Company.

Daddona, P.E., Davidson, B.L., Perignon, J.L., and Kelley, W.N. (1985). Genetic expression in partial adenosine deaminase deficiency. mRNA levels and protein turnover for the enzyme variants in human B-lymphoblast cell lines. J Biol Chem *260*, 3875-3880.

de Souza, S.J. (2003). The emergence of a synthetic theory of intron evolution. Genetica *118*, 117-121.

de Souza, S.J., Long, M., Klein, R.J., Roy, S., Lin, S., and Gilbert, W. (1998). Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. Proc Natl Acad Sci USA *95*, 5094-5099.

Dermitzakis, E.T. and Clark, A.G. (2001). Differential selection after duplication in mammalian developmental genes. Mol Biol Evol *18*, 557-562.

Dolezal, T., Dolezelova, E., Zurovec, M., and Bryant, P.J. (2005). A Role for Adenosine Deaminase in Drosophila Larval Development. PLoS Biol *3*, e201.

Dolezal, T., Gazi, M., Zurovec, M., and Bryant, P.J. (2003). Genetic analysis of the ADGF multigene family by homologous recombination and gene conversion in Drosophila. Genetics *165*, 653-666.

Don, R.H., Cox, P.T., Wainwright, B.J., Baker, K., and Mattick, J.S. (1991). 'Touchdown' PCR to circumvent spurious priming during gene amplification. Nucleic Acids Res *19*, 4008.

Dubey, R.K., Gillespie, D.G., Mi, Z., Suzuki, F., and Jackson, E.K. (1996). Smooth muscle cell-derived adenosine inhibits cell growth. Hypertension *27*, 766-773.

Dubey, R.K., Gillespie, D.G., Mi, Z., and Jackson, E.K. (1997). Exogenous and endogenous adenosine inhibits fetal calf serum-induced growth of rat cardiac fibroblasts: role of A2B receptors. Circulation *96*, 2656-2666.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., Ainscough, R., Almeida, J.P., Babbage, A., Bagguley, C., Bailey, J., Barlow, K., Bates, K.N., Beasley, O., Bird, C.P., Blakey, S., Bridgeman, A.M., Buck, D., Burgess, J., Burrill, W.D., and O'Brien, K.P. (1999). The DNA sequence of human chromosome 22. Nature *402*, 489-495.

Edelmann, L., Pandita, R.K., Spiteri, E., Funke, B., Goldberg, R., Palanisamy, N., Chaganti, R.S., Magenis, E., Shprintzen, R.J., and Morrow, B.E. (1999). A common molecular basis for rearrangement disorders on chromosome 22q11. Hum Mol Genet *8*, 1157-1167.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792-1797.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol *300*, 1005-1016.

Ensenauer, R.E., Adeyinka, A., Flynn, H.C., Michels, V.V., Lindor, N.M., Dawson, D.B., Thorland, E.C., Lorentz, C.P., Goldstein, J.L., McDonald, M.T., Smith, W.E., Simon-Fayard, E., Alexander, A.A., Kulharya, A.S., Ketterling, R.P., Clark, R.D., and Jalal, S.M. (2003). Microduplication 22q11.2, an emerging syndrome: clinical, cytogenetic, and molecular analysis of thirteen patients. Am J Hum Genet *73*, 1027-1040.

Ethier, M.F., Chander, V., and Dobson, J.G.J. (1993). Adenosine stimulates proliferation of human endothelial cells in culture. Am J Physiol *265*, H131-H138.

Fedorov, A., Cao, X., Saxonov, S., de Souza, S.J., Roy, S.W., and Gilbert, W. (2001). Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. Proc Natl Acad Sci USA *98*, 13177-13182.

Fedorova, L. and Fedorov, A. (2003). Introns in gene evolution. Genetica *118*, 123-131.

Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. Nature *340*, 245-246.

Fisher, E. and Scambler, P. (1994). Human haploinsufficiency--one for sorrow, two for joy. Nat Genet *7*, 5-7.

Fleischer, B. (1994). CD26: a surface protease involved in T-cell activation. Immunol Today *15*, 180-184.

Flybase Consortium (1999). The FlyBase database of the Drosophila Genome Projects and community literature. Nucleic Acids Res *27*, 85-88.

Footz, T.K., Brinkman-Mills, P., Banting, G.S., Maier, S.A., Riazi, M.A., Bridgland, L., Hu, S., Birren, B., Minoshima, S., Shimizu, N., Pan, H., Nguyen, T., Fang, F., Fu, Y., Ray, L., Wu, H., Shaull, S., Phan, S., Yao, Z., Chen, F., Huan, A., Hu, P., Wang, Q., Loh, P., Qi, S., Roe, B.A., and McDermid, H.E. (2001). Analysis of the cat eye syndrome critical region in humans and the region of conserved synteny in mice: a search for candidate genes at or near the human chromosome 22 pericentromere. Genome Res *11*, 1053-1070.

Fougerousse, F., Bullen, P., Herasse, M., Lindsay, S., Richard, I., Wilson, D., Suel, L., Durand, M., Robson, S., Abitbol, M., Beckmann, J.S., and Strachan, T. (2000). Human-mouse differences in the embryonic expression patterns of developmental control genes and disease genes. Hum Mol Genet *9*, 165-173.

Fraccaro, M., Lindsten, J., Ford, C.E., and Iselius, L. (1980). The 11q;22q translocation: a European collaborative analysis of 43 cases. Hum Genet *56*, 21-51.

Franco, R., Casado, V., Ciruela, F., Saura, C., Mallol, J., Canela, E.I., and Lluis, C. (1997). Cell surface adenosine deaminase: much more than an ectoenzyme. Prog Neurobiol *52*, 283-294.

Franco, R., Mallol, J., Casado, V., Lluis, C., Canela, E.I., Saura, C., Blanco, J., and Ciruela, F. (1998). Ecto-adenosine deaminase: an ecto-enzyme and a costimulatory protein acting on a variety of cell surface receptors. Drug Development Research *45*, 261-268.

Funke, B., Edelmann, L., McCain, N., Pandita, R.K., Ferreira, J., Merscher, S., Zohouri, M., Cannizzaro, L., Shanske, A., and Morrow, B.E. (1999). Der(22) syndrome and velo-cardio-facial syndrome/DiGeorge syndrome share a 1.5-Mb region of overlap on chromosome 22q11. Am J Hum Genet *64*, 747-758.

Gakis, C. (1996). Adenosine deaminase (ADA) isoenzymes ADA1 and ADA2: diagnostic and biological role. Eur Respir J *9*, 632-633.

Gakis, C., Cappio-Borlino, A., and Pulina, G. (1998). Enzymes (isoenzyme system) as homeostatic mechanisms the isoenzyme (ADA2) of adenosine deaminase of human monocytes-macrophages as a regulator of the 2'deoxyadenosine. Biochem Mol Biol Int *46*, 487-494.

Gamulin, V., Skorokhod, A., Kavsan, V., Muller, I.M., and Muller, W.E. (1997). Experimental indication in favor of the introns-late theory: the receptor tyrosine kinase gene from the sponge Geodia cydonium. J Mol Evol *44*, 242-252.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., and Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res *31*, 3784-3788.

Gibson, C.W., Thomson, N.H., Abrams, W.R., and Kirkham, J. (2005). Nested genes: Biological implications and use of AFM for analysis. Gene *350*, 15-23.

Gilbert, W., Marchionni, M., and McKnight, G. (1986). On the antiquity of introns. Cell *46*, 151-153.

Griffiths, A.J.F., Miller, J.H., Suzuki, D.T., Lewontin, R.C., and Gelbart, W.M. (1996). An introduction to genetic analysis. New York: W.H. Freeman and company.

Gross, M. (1994). Molecular biology of AMP deaminase deficiency. Pharm World Sci *16*, 55-61.

Gu, X. (2001). Maximum-likelihood approach for gene family evolution under functional divergence. Mol Biol Evol *18*, 453-464.

Hasko, G., Kuhel, D.G., Nemeth, Z.H., Mabley, J.G., Stachlewitz, R.F., Virag, L., Lohinai, Z., Southan, G.J., Salzman, A.L., and Szabo, C. (2000). Inosine inhibits inflammatory cytokine production by a post-transcriptional mechanism and protects against endotoxin-induced shock. J Immunol *164*, 1013-1019.

Hernandez, D. and Fisher, E.M. (1999). Mouse autosomal trisomy: two's company, three's a crowd. Trends Genet *15*, 241-247.

Hershfield, M.S. (1995). PEG-ADA replacement therapy for adenosine deaminase deficiency: an update after 8.5 years. Clin Immunol Immunopathol *76*, S228-S232

Hershfield, M.S. (2003). Genotype is an important determinant of phenotype in adenosine deaminase deficiency. Curr Opin Immunol *15*, 571-577.

Hirschhorn, R. and Ratech, H. (1980). Isozymes of adenosine deaminase. Isozymes Curr Top Biol Med Res *4*, 131-157.

Hogan, B., Beddington, R., Costantini, F., and Lacy, E. (1994). Manipulating the Mouse Embryo: A laboratory manual. Plainview, NY: Cold Spring Harbor Laboratory Press.

Holbro, T. and Hynes, N.E. (2004). ErbB receptors: directing key signaling networks throughout life. Annu Rev Pharmacol Toxicol *44*, 195-217.

Holder, M. and Lewis, P.O. (2003). Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet *4*, 275-284.

Homma, K., Matsushita, T., and Natori, S. (1996). Purification, characterization, and cDNA cloning of a novel growth factor from the conditioned medium of NIH-Sape-4, an embryonic cell line of Sarcophaga peregrina (flesh fly). J Biol Chem *271*, 13770-13775.

Homma, K.J., Tanaka, Y., Matsushita, T., Yokoyama, K., Matsui, H., and Natori, S. (2001). Adenosine deaminase activity of insect-derived growth factor is essential for its growth factor activity. J Biol Chem *276*, 43761-43766.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R., and Bollback, J.P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. Science *294*, 2310-2314.

Iwaki-Egawa, S., Namiki, C., and Watanabe, Y. (2004). Adenosine deaminase 2 from chicken liver: purification, characterization, and N-terminal amino acid sequence. Comp Biochem Physiol B Biochem Mol Biol *137*, 247-254.

Jacobson, K.A., Hoffmann, C., Cattabeni, F., and Abbracchio, M.P. (1999). Adenosine-induced cell death: evidence for receptor-mediated signaling. Apoptosis *4*, 197-211.

Johnson, A., Minoshima, S., Asakawa, S., Shimizu, N., Shizuya, H., Roe, B.A., and McDermid, H.E. (1999). A 1.5-Mb contig within the cat eye syndrome critical region at human chromosome 22q11.2. Genomics *57*, 306-309.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci *8*, 275-282.

Karpova, A.Y., Howley, P.M., and Ronco, L.V. (2000). Dual utilization of an acceptor/donor splice site governs the alternative splicing of the IRF-3 gene. Genes Dev *14*, 2813-2818.

Karpova, A.Y., Ronco, L.V., and Howley, P.M. (2001). Functional characterization of interferon regulatory factor 3a (IRF-3a), an alternative splice isoform of IRF-3. Mol Cell Biol *21*, 4169-4176.

Kelly, L.E., Davy, B.E., Berbari, N.F., Robinson, M.L., and El-Hodiri, H.M. (2005). Recombineered Xenopus tropicalis BAC expresses a GFP reporter under the control of Arx transcriptional regulatory elements in transgenic Xenopus laevis embryos. Genesis *41*, 185-191.

Knoll, J.H., Asamoah, A., Pletcher, B.A., and Wagstaff, J. (1995). Interstitial duplication of proximal 22q: phenotypic overlap with cat eye syndrome. Am J Med Genet *55*, 221-224.

Kolaczkowski, B. and Thornton, J.W. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature *431*, 980-984.

Kolls, J.K. and Linden, A. (2004). Interleukin-17 family members and inflammation. Immunity *21*, 467-476.

Kozak, M. (1996). Interpreting cDNA sequences: some insights from studies on translation. Mamm Genome *7*, 563-574.

Krauss, V., Pecyna, M., Kurz, K., and Sass, H. (2005). Phylogenetic mapping of intron positions: a case study of translation initiation factor eIF2gamma. Mol Biol Evol *22*, 74-84.

Landry, J.R., Mager, D.L., and Wilhelm, B.T. (2003). Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet *19*, 640-648.

Lavorgna, G., Dahary, D., Lehner, B., Sorek, R., Sanderson, C.M., and Casari, G. (2004a). In search of antisense. Trends Biochem Sci *29*, 88-94.

Lavorgna, G., Sessa, L., Guffanti, A., Lassandro, L., and Casari, G. (2004b). AntiHunter: searching BLAST output for EST antisense transcripts. Bioinformatics *20*, 583-585.

Lehner, B., Williams, G., Campbell, R.D., and Sanderson, C.M. (2002). Antisense transcripts in the human genome. Trends Genet *18*, 63-65.

Li, L., Krantz, I.D., Deng, Y., Genin, A., Banta, A.B., Collins, C.C., Qi, M., Trask, B.J., Kuo, W.L., Cochran, J., Costa, T., Pierpont, M.E., Rand, E.B., Piccoli, D.A., Hood, L., and Spinner, N.B. (1997). Alagille syndrome is caused by mutations in human Jagged1, which encodes a ligand for Notch1. Nat Genet *16*, 243-251.

Li, S. and Aksoy, S. (2000). A family of genes with growth factor and adenosine deaminase similarity are preferentially expressed in the salivary glands of Glossina m. morsitans. Gene *252*, 83-93.

Lindsay, E.A., Shaffer, L.G., Carrozzo, R., Greenberg, F., and Baldini, A. (1995). De novo tandem duplication of chromosome segment 22q11-q12: clinical, cytogenetic, and molecular characterization. Am J Med Genet *56*, 296-299.

Liu, M. and Grigoriev, A. (2004). Protein domains correlate strongly with exons in multiple eukaryotic genomes--evidence of exon shuffling? Trends Genet *20*, 399-403.

Lutz, B., Kuratani, S., Rugarli, E.I., Wawersik, S., Wong, C., Bieber, F.R., Ballabio, A., and Eichele, G. (1994). Expression of the Kallmann syndrome gene in human fetal brain and in the manipulated chick embryo. Hum Mol Genet *3*, 1717-1723.

Lynch, M. and Katju, V. (2004). The altered evolutionary trajectories of gene duplicates. Trends Genet *20*, 544-549.

Ma, P.F. and Fisher, J.R. (1969). Comparative studies of mammalian adenosine deaminases--some distinctive properties in higher mammals. Comp Biochem Physiol *31*, 771-781.

Maddison, W.P. and Maddison, D.R. (1989). Interactive analysis of phylogeny and character evolution using the computer program MacClade. Folia Primatol *53*, 190-202.

Maier, S.A., Podemski, L., Graham, S.W., McDermid, H.E., and Locke, J. (2001). Characterization of the adenosine deaminase-related growth factor (ADGF) gene family in Drosophila. Gene *280*, 27-36.

Makalowska, I., Lin, C.F., and Makalowski, W. (2005). Overlapping genes in vertebrate genomes. Comput Biol Chem *29*, 1-12.

Matsushita, T., Fujii-Taira, I., Tanaka, Y., Homma, K.J., and Natori, S. (2000). Male-specific IDGF, a novel gene encoding a membrane-bound extracellular signaling molecule expressed exclusively in testis of Drosophila melanogaster. J Biol Chem *275*, 36934-36941.

Mattick, J.S. (1994). Introns: evolution and function. Curr Opin Genet Dev *4*, 823-831.

Mattick, J.S. (2004). The hidden genetic program of complex organisms. Sci Am *291*, 60-67.

McDermid, H.E., Duncan, A.M., Brasch, K.R., Holden, J.J., Magenis, E., Sheehy, R., Burn, J., Kardon, N., Noel, B., and Schinzel, A. (1986). Characterization of the supernumerary chromosome in cat eye syndrome. Science *232*, 646-648.

McDermid, H.E. and Morrow, B.E. (2002). Genomic disorders on 22q11. Am J Hum Genet *70*, 1077-1088.

McTaggart, K.E., Budarf, M.L., Driscoll, D.A., Emanuel, B.S., Ferreira, P., and McDermid, H.E. (1998). Cat eye syndrome chromosome breakpoint clustering: identification of two intervals also associated with 22q11 deletion syndrome breakpoints. Cytogenet Cell Genet *81*, 222-228.

Mears, A.J., Duncan, A.M., Budarf, M.L., Emanuel, B.S., Sellinger, B., Siegel-Bartelt, J., Greenberg, C.R., and McDermid, H.E. (1994). Molecular characterization of the marker chromosome associated with cat eye syndrome. Am J Hum Genet *55*, 134-142.

Mears, A.J., el-Shanti, H., Murray, J.C., McDermid, H.E., and Patil, S.R. (1995). Minute supernumerary ring chromosome 22 associated with cat eye syndrome: further delineation of the critical region. Am J Hum Genet *57*, 667-673.

Meins, M., Burfeind, P., Motsch, S., Trappe, R., Bartmus, D., Langer, S., Speicher, M.R., Muhlendyck, H., Bartels, I., and Zoll, B. (2003). Partial trisomy of chromosome 22 resulting from an interstitial duplication of 22q11.2 in a child with typical cat eye syndrome. J Med Genet *40*, e62.

Meng, J.P., Zhang, F.P., Huhtaniemi, I., and Pakarinen, P. (1997). Characterization and developmental expression of a testis-specific adenosine deaminase mRNA in the mouse. J Androl *18*, 88-95.

Migchielsen, A.A., Breuer, M.L., van Roon, M.A., te Riele, H., Zurcher, C., Ossendorp, F., Toutain, S., Hershfield, M.S., Berns, A., and Valerio, D. (1995). Adenosine-deaminase-deficient mice die perinatally and exhibit liver-cell degeneration, atelectasis and small intestinal cell death. Nat Genet *10*, 279-287.

Modrek, B. and Lee, C.J. (2003). Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet *34*, 177-180.

Mohamedali, K.A., Kurz, L.C., and Rudolph, F.B. (1996). Site-directed mutagenesis of active site glutamate-217 in mouse adenosine deaminase. Biochem *35*, 1672-1680.

Morton, N.E. (1991). Parameters of the human genome. Proc Natl Acad Sci USA *88*, 7474-7476.

Mousseau, I. (2005). CECR6: Evidence of two overlapping reading frames in the cat eye syndrome critical region. MSc thesis. University of Alberta.

Mozdziak, P.E. and Petitte, J.N. (2004). Status of transgenic chicken models for developmental biology. Dev Dyn *229*, 414-421.

Muesch, A., Hartmann, E., Rohde, K., Rubartelli, A., Sitia, R., and Rapoport, T.A. (1990). A novel pathway for secretory proteins? Trends Biochem Sci *15*, 86-88.

Nasevicius, A. and Ekker, S.C. (2000). Effective targeted gene 'knockdown' in zebrafish. Nat Genet *26*, 216-220.

Niedzwicki, J.G. and Abernethy, D.R. (1991). Structure-activity relationship of ligands of human plasma adenosine deaminase2. Biochem Pharmacol *41*, 1615-1624.

Niedzwicki, J.G., Liou, C., Abernethy, D.R., Lima, J.E., Hoyt, A., Lieberman, M., and Bethlenfalvay, N.C. (1995). Adenosine deaminase isoenzymes of the opossum Didelphis virginiana: initial chromatographic and kinetic studies. Comp Biochem Physiol B Biochem Mol Biol *111*, 291-298.

Oda, T., Elkahloun, A.G., Pike, B.L., Okajima, K., Krantz, I.D., Genin, A., Piccoli, D.A., Meltzer, P.S., Spinner, N.B., Collins, F.S., and Chandrasekharappa, S.C. (1997). Mutations in the human Jagged1 gene are responsible for Alagille syndrome. Nat Genet *16*, 235-242.

Oyler, M., Long, B.W., and Cox, L.A. (2004). Sonographic markers used to detect frequent trisomies. Radiol Technol *76*, 13-18.

Page, R.D. (1996). TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci *12*, 357-358.

Painter, S.D., Kalman, V.K., Nagle, G.T., Zuckerman, R.A., and Blankenship, J.E. (1985). The anatomy and functional morphology of the large hermaphroditic duct of three species of Aplysia, with special reference to the atrial gland. J Morphol *186*, 167-194.

Papadakis, M.N. and Patrinos, G.P. (1999). Contribution of gene conversion in the evolution of the human beta-like globin gene family. Hum Genet *104*, 117-125.

Ramalho-Santos, M., Melton, D.A., and McMahon, A.P. (2000). Hedgehog signals regulate multiple aspects of gastrointestinal development. Development *127*, 2763-2772.

Reiss, J.A., Weleber, R.G., Brown, M.G., Bangs, C.D., Lovrien, E.W., and Magenis, R.E. (1985). Tandem duplication of proximal 22q: a cause of cat-eye syndrome. Am J Med Genet *20*, 165-171.

Riazi, M.A., Brinkman-Mills, P., Nguyen, T., Pan, H., Phan, S., Ying, F., Roe, B.A., Tochigi, J., Shimizu, Y., Minoshima, S., Shimizu, N., Buchwald, M., and

McDermid, H.E. (2000). The human homologue of insect-derived growth factor, CECR1, is a candidate gene for features of cat eye syndrome. Genomics *64*, 277-285.

Ribard, C., Rochet, M., Labedan, B., Daignan-Fornier, B., Alzari, P., Scazzocchio, C., and Oestreicher, N. (2003). Sub-families of alpha/beta barrel enzymes: a new adenine deaminase family. J Mol Biol *334*, 1117-1131.

Ribeiro, J.M., Charlab, R., and Valenzuela, J.G. (2001). The salivary adenosine deaminase activity of the mosquitoes Culex quinquefasciatus and Aedes aegypti. J Exp Biol *204*, 2001-2010.

Riveros-Rosas, H., Julian-Sanchez, A., Villalobos-Molina, R., Pardo, J.P., and Pina, E. (2003). Diversity, taxonomy and evolution of medium-chain dehydrogenase / reductase superfamily. Eur J Biochem *270*, 3309-3334.

Rogozin, I.B., Lyons-Weiler, J., and Koonin, E.V. (2000). Intron sliding in conserved gene families. Trends Genet *16*, 430-432.

Ronquist, F. and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics *19*, 1572-1574.

Rosias, P.R., Sijstermans, J.M., Theunissen, P.M., Pulles-Heintzberger, C.F., De Die-Smulders, C.E., Engelen, J.J., and Van Der Meer, S.B. (2001). Phenotypic variability of the cat eye syndrome. Case report and review of the literature. Genet Couns *12*, 273-282.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor, M.G., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., Cherry, J.M., Henikoff, S., Skupski, M.P., Misra, S., Ashburner, M., Birney, E., Boguski, M.S., Brody, T., Brokstein, P., Celniker, S.E., Chervitz, S.A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R.F., Gelbart, W.M., George, R.A., Goldstein, L.S., Gong, F., Guan, P., Harris, N.L., Hay, B.A., Hoskins, R.A., Li, J., Li, Z., Hynes, R.O., Jones, S.J., Kuehl, P.M., Lemaitre, B., Littleton, J.T., Morrison, D.K., Mungall, C., O'Farrell, P.H., Pickeral, O.K., Shue, C., Vosshall, L.B., Zhang, J., Zhao, Q., Zheng, X.H., and Lewis, S. (2000). Comparative genomics of the eukaryotes. Science *287*, 2204-2215.

Rzhetsky, A., Ayala, F.J., Hsu, L.C., Chang, C., and Yoshida, A. (1997). Exon/intron structure of aldehyde dehydrogenase genes supports the "introns-late" theory. Proc Natl Acad Sci USA *94*, 6820-6825.

Sambrook, J. and Russell, D.W. (2001). Molecular Cloning: A laboratory manual. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Schinzel, A., Schmid, W., Fraccaro, M., Tiepolo, L., Zuffardi, O., Opitz, J.M., Lindsten, J., Zetterqvist, P., Enell, H., Baccichetti, C., Tenconi, R., and Pagon, R.A. (1981). The "cat eye syndrome": dicentric small marker chromosome probably derived from a no.22 (tetrasomy 22pter to q11) associated with a characteristic phenotype. Report of 11 patients and delineation of the clinical picture. Hum Genet *57*, 148-158.

Schmitt, D.M. and Brower, D.L. (2001). Intron dynamics and the evolution of integrin beta-subunit genes: maintenance of an ancestral gene structure in the coral, Acropora millepora. J Mol Evol *53*, 703-710.

Shaikh, T.H., Budarf, M.L., Celle, L., Zackai, E.H., and Emanuel, B.S. (1999). Clustered 11q23 and 22q11 breakpoints and 3:1 meiotic malsegregation in multiple unrelated t(11;22) families. Am J Hum Genet *65*, 1595-1607.

Shearwin, K.E., Callen, B.P., and Egan, J.B. (2005). Transcriptional interference--a crash course. Trends Genet *21*, 339-345.

Sideraki, V., Wilson, D.K., Kurz, L.C., Quiocho, F.A., and Rudolph, F.B. (1996). Site-directed mutagenesis of histidine 238 in mouse adenosine deaminase: substitution of histidine 238 does not impede hydroxylate formation. Biochemistry *35*, 15019-15028.

Sossin, W., Kreiner, T., Barinaga, M., Schilling, J., and Scheller, R. (1989). A Dense Core Vesicle Protein Is Restricted to the Cortex of Granules in the Exocrine Atrial Gland of Aplysia californica. J Biol Chem *264*, 16933-16940.

Sterne, G.D., Coulton, G.R., Brown, R.A., Green, C.J., and Terenghi, G. (1997). Neurotrophin-3-enhanced nerve regeneration selectively improves recovery of muscle fibers expressing myosin heavy chains 2b. J Cell Biol *139*, 709-715.

Stoltzfus, A., Logsdon, J.M.J., Palmer, J.D., and Doolittle, W.F. (1997). Intron "sliding" and the diversity of intron positions. Proc Natl Acad Sci USA *94*, 10739-10744.

Summerton, J. and Weller, D. (1997). Morpholino antisense oligomers: design, preparation, and properties. Antisense Nucleic Acid Drug Dev *7*, 187-195.

Swofford, D. L. (2001). PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods), version 4.0b8. Sunderland, MA, Sinauer Associates Inc.

Tassabehji, M. (2003). Williams-Beuren syndrome: a challenge for genotype-phenotype correlations. Hum Mol Genet *12 Spec No 2*, R229-R237.

Taylor, J.S., Braasch, I., Frickey, T., Meyer, A., and Van de Peer, Y. (2003). Genome duplication, a trait shared by 22000 species of ray-finned fish. Genome Res *13*, 382-390.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res *22*, 4673-4680.

Tyshenko, M.G. and Walker, V.K. (1997). Towards a reconciliation of the introns early or late views: triosephosphate isomerase genes from insects. Biochim Biophys Acta *1353*, 131-136.

Ungerer, J.P., Oosthuizen, H.M., Bissbort, S.H., and Vermaak, W.J. (1992). Serum adenosine deaminase: isoenzymes and diagnostic application. Clin Chem *38*, 1322-1326.

Van der Weyden, M.B. and Kelley, W.N. (1976). Human adenosine deaminase. Distribution and properties. J Biol Chem *251*, 5448-5456.

Venkatesh, B., Ning, Y., and Brenner, S. (1999). Late changes in spliceosomal introns define clades in vertebrate evolution. Proc Natl Acad Sci USA *96*, 10267-10271.

Voiculescu, O., Taillebourg, E., Pujades, C., Kress, C., Buart, S., Charnay, P., and Schneider-Maunoury, S. (2001). Hindbrain patterning: Krox20 couples segmentation and specification of regional identity. Development *128*, 4967-4978.

Wakamiya, M., Blackburn, M.R., Jurecic, R., McArthur, M.J., Geske, R.S., Cartwright, J.J., Mitani, K., Vaishnav, S., Belmont, J.W., and Kellems, R.E. (1995). Disruption of the adenosine deaminase gene causes hepatocellular impairment and perinatal lethality in mice. Proc Natl Acad Sci USA *92*, 3673-3677.

Wang, Z. and Quiocho, F.A. (1998). Complexes of adenosine deaminase with two potent inhibitors: X-ray structures in four independent molecules at pH of maximum activity. Biochemistry *37*, 8314-8324.

Watkins, B.P., Bolender, D.L., Lough, J., and Kolesari, G.L. (1998). Teratogenic effects of implanting fibroblast growth factor-2-soaked beads in the cardiac region of the stage 24 chick embryo. Teratology *57*, 140-145.

Weijer, C.J. (2004). Dictyostelium morphogenesis. Curr Opin Genet Dev *14*, 392-398.

Westneat, D.F., Noon, W.A., Reeve, H.K., and Aquadro, C.F. (1988). Improved hybridization conditions for DNA 'fingerprints' probed with M13. Nucleic Acids Res *16*, 4161

Wilkinson, D.G. and Nieto, M.A. (1993). Detection of messenger RNA by in situ hybridization to tissue sections and whole mounts. Methods Enzymol *225*, 361-373.

Wilson, D.K., Rudolph, F.B., and Quiocho, F.A. (1991). Atomic structure of adenosine deaminase complexed with a transition-state analog: understanding catalysis and immunodeficiency mutations. Science *252*, 1278-1284.

Witowski, J., Ksiazek, K., and Jorres, A. (2004). Interleukin-17: a mediator of inflammatory responses. Cell Mol Life Sci *61*, 567-579.

Yagi, H., Furutani, Y., Hamada, H., Sasaki, T., Asakawa, S., Minoshima, S., Ichida, F., Joo, K., Kimura, M., Imamura, S., Kamatani, N., Momma, K., Takao, A., Nakazawa, M., Shimizu, N., and Matsuoka, R. (2003). Role of TBX1 in human del22q11.2 syndrome. Lancet *362*, 1366-1373.

Yao, Z., Painter, S.L., Fanslow, W.C., Ulrich, D., Macduff, B.M., Spriggs, M.K., and Armitage, R.J. (1995). Human IL-17: a novel cytokine derived from T-cells. J Immunol *155*, 5483-5486.

Yao, Z., Spriggs, M.K., Derry, J.M., Strockbine, L., Park, L.S., VandenBos, T., Zappone, J.D., Painter, S.L., and Armitage, R.J. (1997). Molecular characterization of the human interleukin (IL)-17 receptor. Cytokine *9*, 794-800.

Young, K.H. (1998). Yeast two-hybrid: so many interactions, (in) so little time. Biol Reprod *58*, 302-311.

Zackai, E.H. and Emanuel, B.S. (1980). Site-specific reciprocal translocation, t(11;22) (q23;q11), in several unrelated families with 3:1 meiotic disjunction. Am J Med Genet *7*, 507-521.

Zavialov, A.V. and Engstrom, A. (2005). Human ADA2 belongs to a new family of growth factors with adenosine deaminase activity. Biochem J, *in press*.

Zurovec, M., Dolezal, T., Gazi, M., Pavlova, E., and Bryant, P.J. (2002). Adenosine deaminase-related growth factors stimulate cell proliferation in Drosophila by depleting extracellular adenosine. Proc Natl Acad Sci USA *99*, 4403-4408.

Appendix

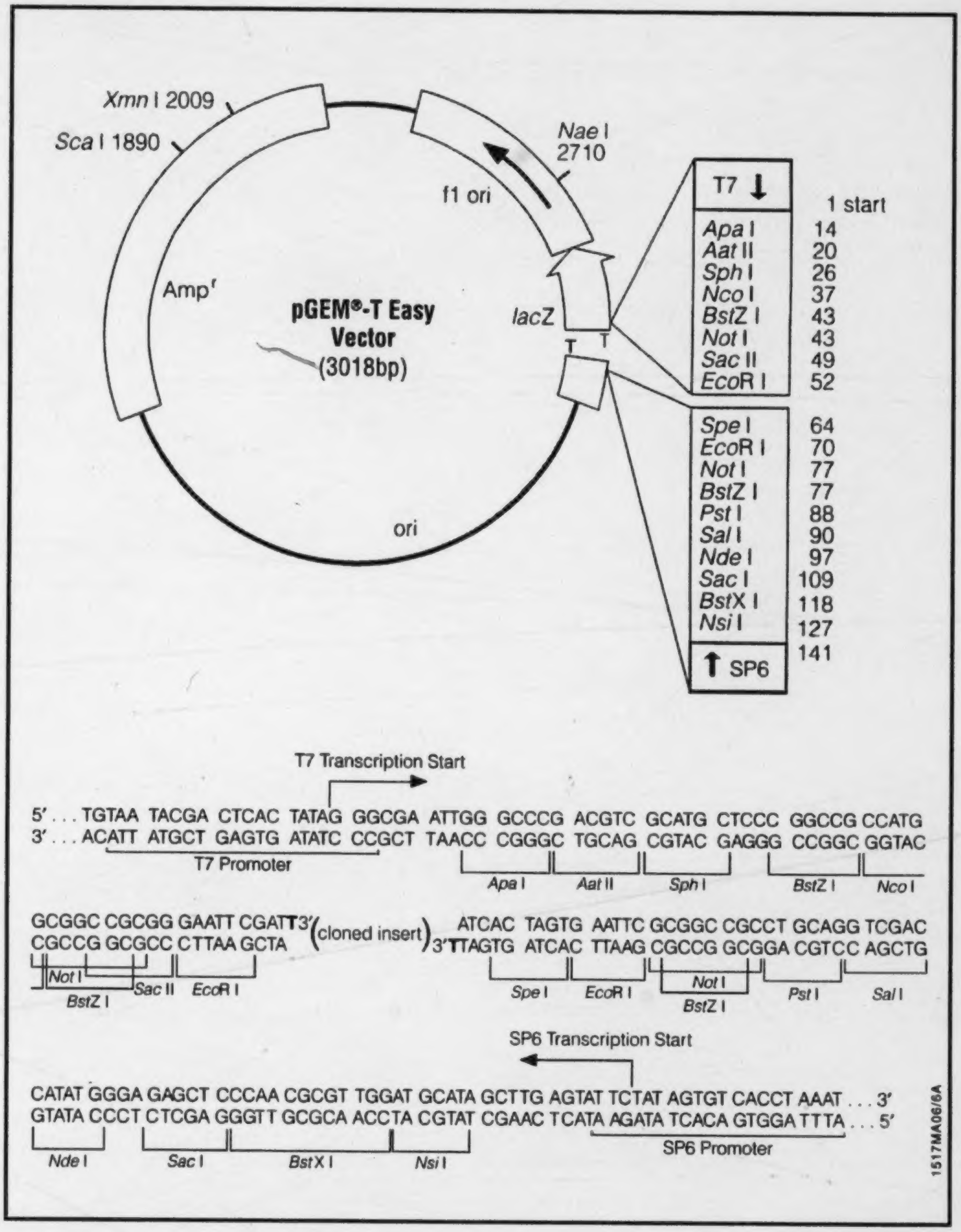
Xmn I 2009
Sca I 1890
Nae I
2710
f1 ori
Ampʳ
pGEM®-T Easy
Vector
(3018bp)
lacZ
T
T
ori
T7 ↓
1 start
Apa I 14
Aat II 20
Sph I 26
Nco I 37
BstZ I 43
Not I 43
Sac II 49
EcoR I 52
Spe I 64
EcoR I 70
Not I 77
BstZ I 77
Pst I 88
Sal I 90
Nde I 97
Sac I 109
BstX I 118
Nsi I 127
141
↑ SP6
T7 Transcription Start
5′ . . . TGTAA TACGA CTCAC TATAG GGCGA ATTGG GCCCG ACGTC GCATG CTCCC GGCCG CCATG
3′ . . . ACATT ATGCT GAGTG ATATC CCGCT TAACC CGGGC TGCAG CGTAC GAGGG CCGGC GGTAC
T7 Promoter
Apa I
Aat II
Sph I
BstZ I
Nco I
GCGGC CGCGG GAATT CGATT 3′ (cloned insert) ATCAC TAGTG AATTC GCGGC CGCCT GCAGG TCGAC
CGCCG GCGCC CTTAA GCTA 3′TTAGTG ATCAC TTAAG CGCCG GCGGA CGTCC AGCTG
Not I
BstZ I
Sac II
EcoR I
Spe I
EcoR I
Not I
BstZ I
Pst I
Sal I
SP6 Transcription Start
CATAT GGGA GAGCT CCCAA CGCGT TGGAT GCATA GCTTG AGTAT TCTAT AGTGT CACCT AAAT . . . 3′
GTATA CCCT CTCGA GGGTT GCGCA ACCTA CGTAT CGAAC TCATA AGATA TCACA GTGGA TTTA . . . 5′
Nde I
Sac I
BstX I
Nsi I
SP6 Promoter
1517MA06/6A

Figure A1. Vector map and multiple cloning site of the pGEM-T Easy vector (Promega) used for cloning PCR products.

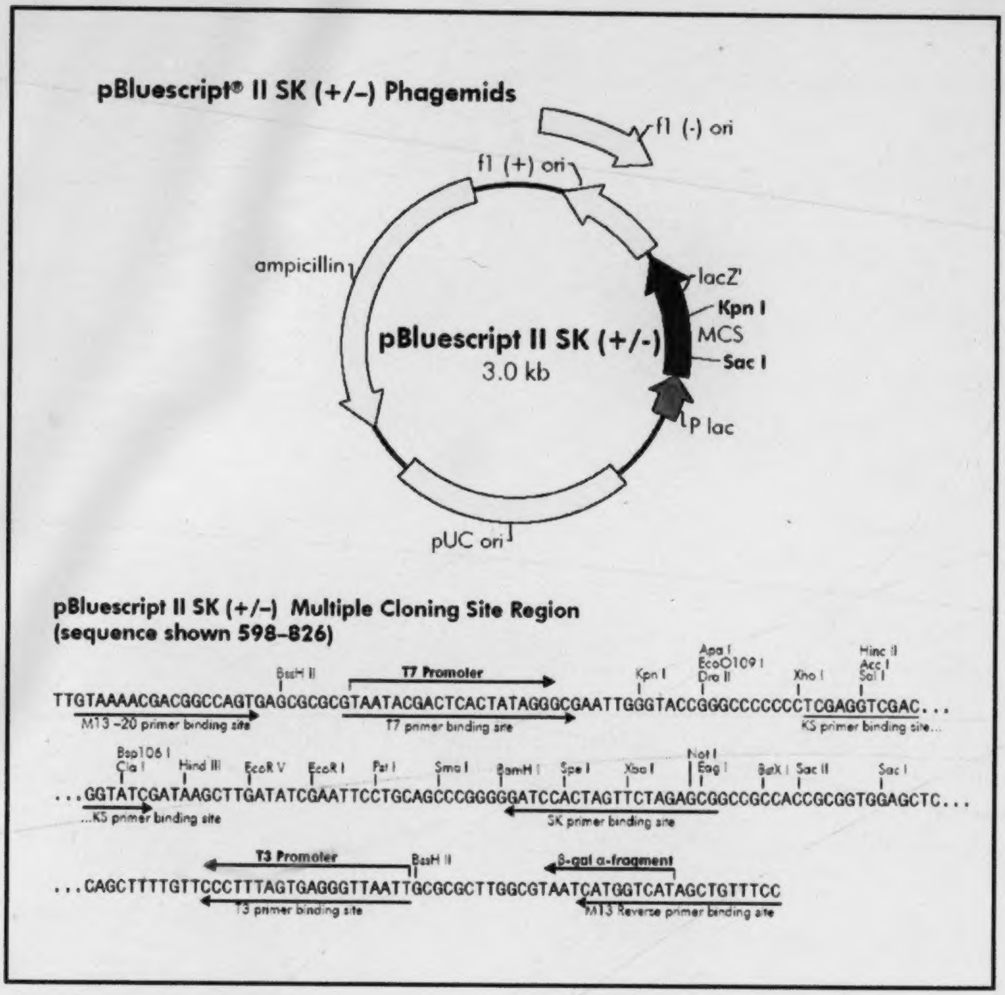
pBluescript® II SK (+/–) Phagemids
f1 (-) ori
f1 (+) ori
ampicillin
lacZ'
Kpn I
MCS
Sac I
pBluescript II SK (+/-)
3.0 kb
P lac
pUC ori
pBluescript II SK (+/–) Multiple Cloning Site Region
(sequence shown 598–826)
BssH II
T7 Promoter
Kpn I
Apa I
EcoO109 I
Dra II
Xho I
Hinc II
Acc I
Sal I
TTGTAAAACGACGGCCAGTGAGCGCGCGTAATACGACTCACTATAGGGCGAATTGGGTACCGGGCCCCCCCTCGAGGTCGAC...
M13 –20 primer binding site
T7 primer binding site
KS primer binding site...
Bsp106 I
Cla I
Hind III
EcoR V
EcoR I
Pst I
Sma I
BamH I
Spe I
Xba I
Not I
Eag I
BstX I
Sac II
Sac I
...GGTATCGATAAGCTTGATATCGAATTCCTGCAGCCCGGGGGATCCACTAGTTCTAGAGCGGCCGCCACCGCGGTGGAGCTC...
...KS primer binding site
SK primer binding site
T3 Promoter
BssH II
β-gal α-fragment
...CAGCTTTTGTTCCCTTTAGTGAGGGTTAATTGCGCGCTTGGCGTAATCATGGTCATAGCTGTTTCC
T3 primer binding site
M13 Reverse primer binding site

Figure A2. Vector map and multiple cloning site of pBluescript II SK- (Stratagene).

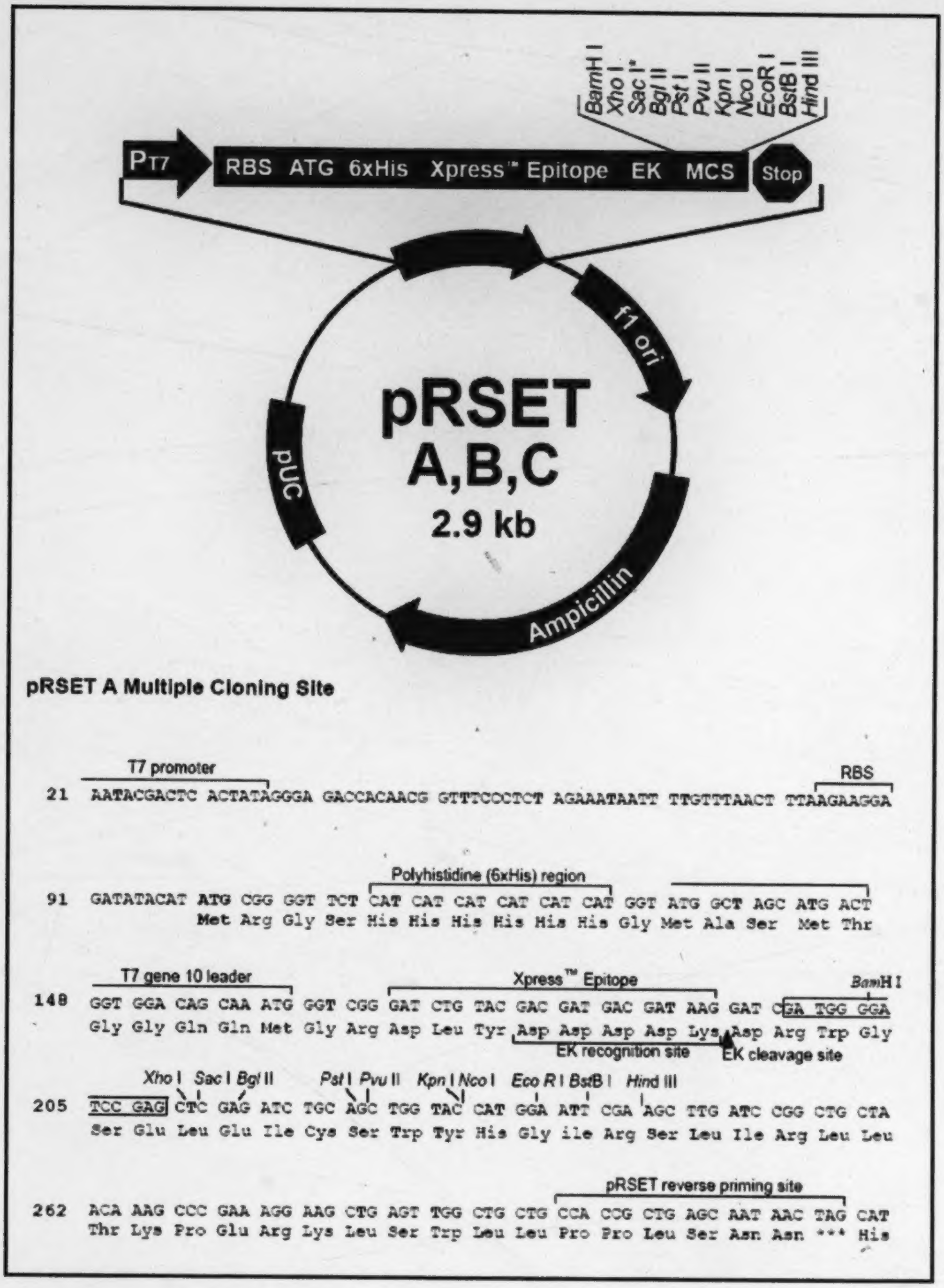
BamH I
Xho I
Sac I*
Bgl II
Pst I
Pvu II
Kpn I
Nco I
EcoR I
BstB I
Hind III
PT7
RBS ATG 6xHis Xpress™ Epitope EK MCS
Stop
f1 ori
pRSET
A,B,C
2.9 kb
pUC
Ampicillin
pRSET A Multiple Cloning Site
T7 promoter
RBS
21 AATACGACTC ACTATAGGGA GACCACAACG GTTTCCCTCT AGAAATAATT TTGTTTAACT TTAAGAAGGA
Polyhistidine (6xHis) region
91 GATATACAT ATG CGG GGT TCT CAT CAT CAT CAT CAT CAT GGT ATG GCT AGC ATG ACT
Met Arg Gly Ser His His His His His His Gly Met Ala Ser Met Thr
T7 gene 10 leader
Xpress™ Epitope
BamH I
148 GGT GGA CAG CAA ATG GGT CGG GAT CTG TAC GAC GAT GAC GAT AAG GAT CGA TGG GGA
Gly Gly Gln Gln Met Gly Arg Asp Leu Tyr Asp Asp Asp Asp Lys Asp Arg Trp Gly
EK recognition site
EK cleavage site
Xho I
Sac I
Bgl II
Pst I
Pvu II
Kpn I
Nco I
Eco R I
BstB I
Hind III
205 TCC GAG CTC GAG ATC TGC AGC TGG TAC CAT GGA ATT CGA AGC TTG ATC CGG CTG CTA
Ser Glu Leu Glu Ile Cys Ser Trp Tyr His Gly ile Arg Ser Leu Ile Arg Leu Leu
pRSET reverse priming site
262 ACA AAG CCC GAA AGG AAG CTG AGT TGG CTG CTG CCA CCG CTG AGC AAT AAC TAG CAT
Thr Lys Pro Glu Arg Lys Leu Ser Trp Leu Leu Pro Pro Leu Ser Asn Asn *** His

Figure A3. Vector map and multiple cloning site for the pRSET A (Invitrogen) expression vector.

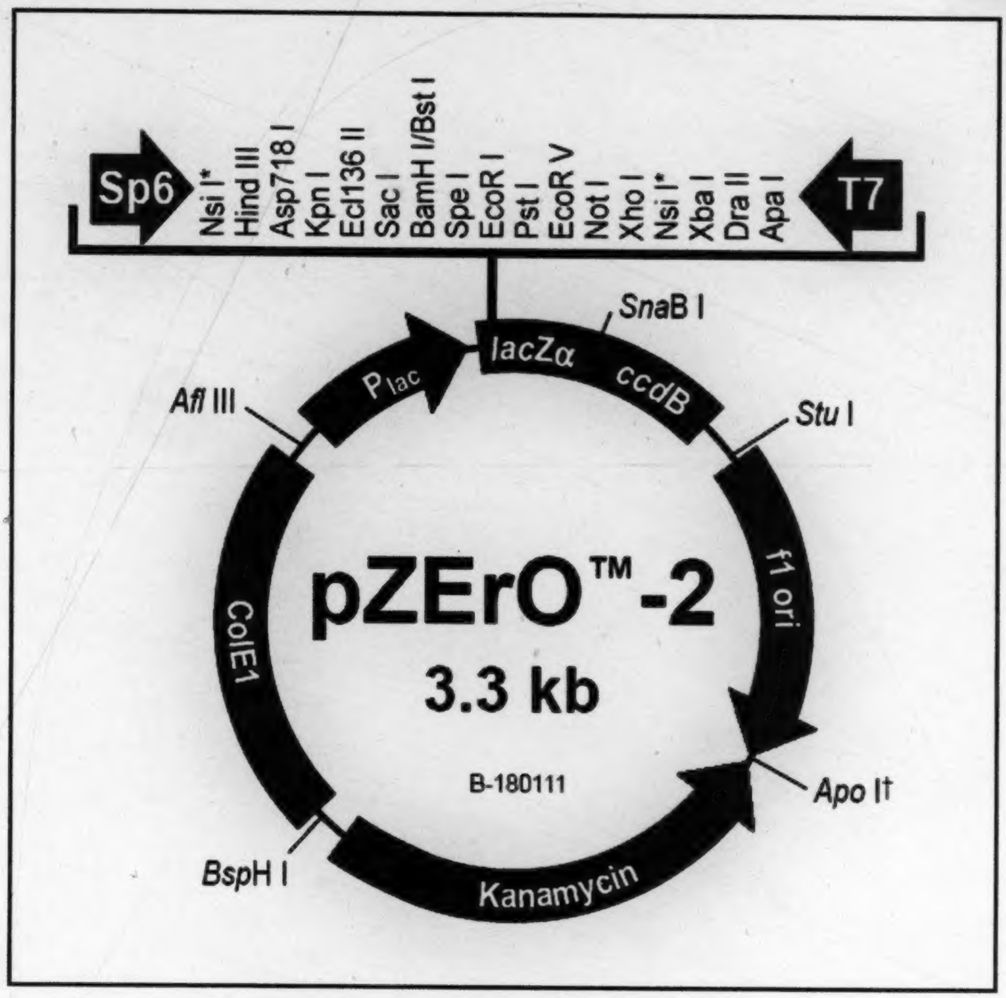
Sp6
Nsi I*
Hind III
Asp718 I
Kpn I
Ecl136 II
Sac I
BamH I/Bst I
Spe I
EcoR I
Pst I
EcoR V
Not I
Xho I
Nsi I*
Xba I
Dra II
Apa I
T7
SnaB I
lacZα
ccdB
Plac
Afl III
Stu I
f1 ori
ColE1
pZErO™-2
3.3 kb
B-180111
Apo I†
BspH I
Kanamycin

Figure A4. Vector map of pZErO-2 (Invitrogen).